АНАЛИЗ ТЕХНОЛОГИЙ РАСПОЗНАВАНИЯ ТЕКСТА ИЗ ИЗОБРАЖЕНИЯ

К.А. Бобров, *магистр*
В.Д. Шульман, *магистр*
К.П. Власов, *магистр*
Московский государственный технический университет им. Н.Э. Баумана (Россия, г. Москва)

DOI:10.24412/2500-1000-2022-3-2-124-128

Аннотация. *В статье анализируются различные технологии для распознавания текста из изображения. В рамках статьи рассматриваются шаги алгоритма распознавания текста из изображения. Описываются основные методы алгоритма распознавания символов (OCR), приводятся их достоинства и недостатки. По результатам сравнения, получено, что нейросетевой метод распознавания является самым перспективным. Делается анализ библиотек, которые активно используют внутри себя данный нейросетевой метод. В заключении делается вывод об целесообразности использования рассмотренных библиотек в зависимости от условий и специфик задач.*

Ключевые слова: *распознавание символов, сегментация, классификация изображений, постобработка, нейросетевой метод, технология Tesseract.*

В настоящее время информационные технологии находятся на пике развития, создаются и развиваются новые технологии, а также направления. Одно из таких направлений – это распознавание текстов из изображений OCR. Данное направление очень часто встречается в жизни людей. Оно позволяет быстро и точно распознать текстовую информацию с фотографий и преобразовать в необходимый формат, будь то текст в консоли или же отдельный файл с выбранным расширением.

Алгоритм распознавания текста из изображения. Распознавание текста из изображения или же оптическое распознавание символов [1] (англ. optical character recognition, OCR) – это технология для автоматизации извлечения данных из печатного, письменного текста, отсканированного документа, файла изображения с целью последующего преобразования текста в машиночитаемую форму. Данная форма будет использована для работы с данными, например, редактирование или поиск информации.

Каждая система OCR состоит из одних и тех же шагов алгоритма:

- предобработка;
- сегментация;
- выделение признаков;
- распознавание символов или классификация;
- постобработка и исправления ошибок распознавания.

Эти алгоритмические шаги выполняются последовательно, и каждый результат шага подается на вход следующего шага. На рисунке 1 представлена схема алгоритма системы распознавания символов.

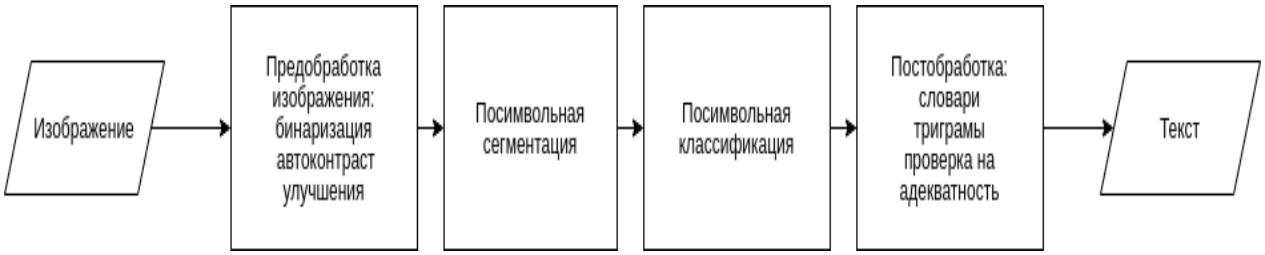

Рис. 1. последовательность работы алгоритма распознавания текста

Шаг предобработки. Перед тем как передать изображение на распознавание, необходимо его обработать и выделить необходимую информацию, как раз для этого и используется слой предобработки. На этом этапе с изображением могут происходить операции очистки изображения от шумов, приведение к виду, позволяющему выделить символы на фоне, фильтрация изображения, сглаживание и увеличение контрастности. Если текст рукописный, то дополнительно применяют подход по выпрямлению символов, так как многие пишут символы с наклоном. В основном используется бинаризация [2] изображения, которая позволяет точно выделить текст и убрать фон.

Алгоритм сегментации. Сегментация изображения [3] – это выделение полезной информации из изображения, с последующей ее обработкой. Сегментация в области распознавания текста состоит из нескольких этапов:

- сегментация строк – выделяем на изображении линиями фрагменты слов;
- сегментация слов – выделение слов, выделяем отдельный фрагменты изображений, где присутствуют слова;
- сегментация символов – разделяется распознанное изображение слова на символы.

Шаг классификации изображения. Классификация позволяет распознать символ из изображения и перевести его в машиночитаемый формат. Существуют разные виды алгоритмов распознавания [4], самыми популярными являются:

- шаблонные алгоритмы;
- признаковые алгоритмы;
- нейросетевые алгоритмы.

Шаблонные алгоритмы. Суть метода заключается в том, что идет сравнение каждого символа с шаблонами из базы. Наиболее подходящим шаблоном считается тот, у которого будет наименьшее количество точек, отличных от исследуемого изображения. Шаблоны для каждого символа обычно получаются усреднением изображений символов обучающей выборки. У данного алгоритма высокая точность распознавания текста, а недостатком является то, что нельзя распознать другой шрифт, который отличается от заложенного в систему. Данный метод должен заранее знать шрифт, который он распознает, именно этот момент ограничивает универсальность шаблонных алгоритмов.

Признаковые алгоритмы. Признаковый метод состоит в том, что изображение представляется как N-мерный вектор признаков. Распознавание заключается в сравнение его с набором эталонных векторов той же размерности. Принятие решение о схожести образа к определенному символу строится на основании математических решений в рамках детерминистического и вероятностного подходов. В системе распознавания данного метода используется классификация, основанная на подсчете евклидового расстояния между вектором признаков распознаваемого символа и векторами признаков эталонного описания. Количество и тип признаков могут определить качество распознавания. Создание вектора происходит во время анализа изображения, такой процесс называют извлечением признаков. Эталонные векторы для символов получают аналогичной обработкой символов обучающей выборки.

Главные достоинства признаковых методов – это простота их реализации, хорошая устойчивость к изменениям формы символов, низкое количество ошибок при распознавании, высокое быстродействие. Самые главные недостатки данного алгоритма – неустойчивость к различным дефектам изображения, например шум, а также на этапе извлечения признаков из символа происходит потеря основной информации, извлечение ведется независимо, из-за чего расположение элементов символа утрачивается.

Нейросетевые алгоритмы. С развитием машинного обучения, а также нейронных сетей, все чаще для распознавания символов используют алгоритмы, построенные с помощью нейронных сетей глубокого обучения. Существует много моделей классификаторов распознавания текста, но всегда в качестве базовых архитектур используются сверточные нейронные сети, а также функционал сохранения и накопления результата распознавания, и рекурентная сеть для распознавания.

Входными данными для нейросетевого метода являются изображения строк и слов. Выходными данными – символы, идущие по порядку, формирующие машинный текст.

Примерная модель классификатора представлена на рисунке 2.

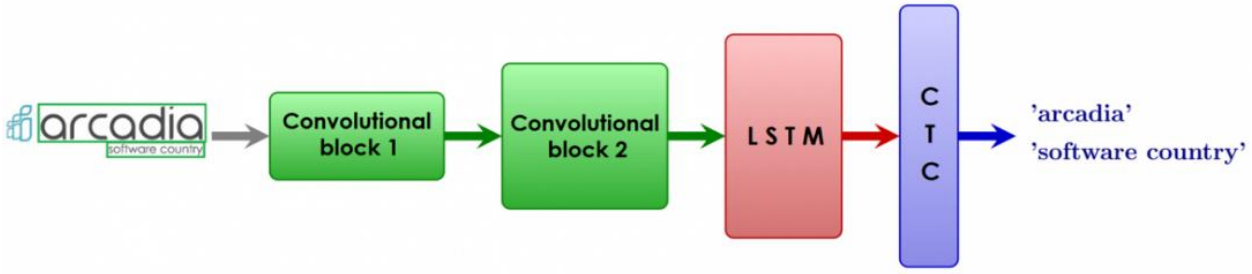

Рис. 2. Архитектура нейросетевого классификатора

На последнем этапе используется слой CTC, который построен на базе нейронной сети, для решения проблем последовательности, основная его задача в OCR – это сохранение последовательности вводимых символов.

Основные недостатки – текст должен быть в вертикальном положении, сложность подбора обучающей выборки. Основные достоинства – это высокая скорость и обобщённость. Именно поэтому данный метод сейчас используется в разных современных системах распознавания текста.

Алгоритм постобработки. Во многих системах OCR результат, получаемый после классификации, не считается достаточным. Необходимо использовать контекстную информацию, которая позволяет не только находить ошибки, но и исправлять их. Существуют разные методы осуществления постобработки, например, глобальные и локальные позиционные диаграммы, триграммы, n-граммы, словари и различные сочетания всех этих методов. Самым популярным подходом является словарь.

Библиотека Tesseract. Библиотека Tesseract [5] бесплатна и просто в использовании. Она несет в себе функциональность инструмента командной строки, но есть также и оболочка для языка Python, которая называется pytesseract, а также приложение для компьютеров с графическим интерфейсом gImageReader. Библиотека Tesseract OCR довольно хорошо распознает отсканированный текст, но, когда дело доходит до рукописного текста, процент распознанного текста снижается и появляются ошибки. С распознаванием табличной информации у Tesseract OCR имеются трудности, необходимо самостоятельно обрабатывать выходные данные с помощью дополнительных технологий и библиотек.

Продукт ABBYY FineReader. Продукт ABBYY FineReader [6] – разработка компании ABBYY, которая входит в число ведущих компаний по распознаванию текста. Данный продукт представляет представляет собой программу с графическим интерфейсом пользователя, где можно загружать документы и получать результат в виде файла. Также существует ABBYY Cloud OCR SDK API – это облачный сервис, который использует движок ABBYY FineReader OCR. В отличие от Tesseract, ABBYY Cloud OCR платный. ABBYY FineReader не имеет проблем с хорошо отсканированным текстом и неплохо справляется с документами, которые сфотографированы и, возможно, с каким-то шумами и разворотами. Однако в рукописном документе он полностью не работает. Его главное достоинство - возможность извлечения таблицы. Помимо ячеек, он извлекает такие мелкие детали как шрифты.

Продукт Google Cloud Vision. Продукт Google Cloud Vision [7], представляет из себя облачный сервис по распознаванию текстов из изображений. Он также, как и продукт ABBYY является платным. Google хорошо справляется с отсканированным текстом и распознает текст в до-

кументе, снятом на камеру, так же, как и ABBYY. Однако он намного лучше, чем Tesseract или ABBYY в распознавании почерка. Google Cloud Vision не очень хорошо обрабатывает таблицы: он извлекает текст, но это все. Фактически, результат работы Cloud Vision представляет собой файл JSON, содержащий информацию о позициях символов. Как и в случае с Tesseract, на основе этой информации можно попытаться обнаружить таблицы, но эта функция не встроена и необходимо задействовать дополнительные ресурсы и технологии.

Обобщенные отличия технологий OCR представлены в таблице 1.

Таблица 1. Анализ OCR технологий

OCR системы	Распознание Отсканированного документа	Распознавание рукописного текста	Распознавание сфотографированного текста	Распознавание таблицы
Tesseract	Хорошо	Плохо	Приемлемо	Плохо, необходимо использовать дополнительные библиотеки
ABBYY FineReader	Хорошо	Плохо	Хорошо	Хорошо
Google Cloud Vision	Хорошо	Приемлемо, имеются ошибки в распознавании	Хорошо	Плохо, доп. библиотеки

Заключение. Выбор той или иной технологии для OCR распознавания зависит от задачи. Для отсканированного документа может подойти Tesseract OCR, он бесплатен и довольно хорошо справляется с поставленной задачей, также можно использовать коммерческие продукты от разных компаний. Для распознавания рукописного текста продукт Google Cloud Vision отлично подходит, так как является одним из жизнеспособных вариантов на сегодня. Если качество документа плохое или же оно сфотографировано с какими-либо шумами и дефектами, то ABBYY FineReader и Google Cloud Vision распознают такой текст. С задачей извлечения табличной информации хорошо работает ABBYY FineReader, который может сохранять такие мелочи как тип и размер шрифта.

Библиографический список

1. Оптическое распознавание символов (OCR). – [Электронный ресурс]. – Режим доступа: http://wiki.technicalvision.ru/index.php/%D0%9E%D (дата обращения: 19.03.2022)

2. Бинаризация изображений: алгоритм Брэдли. – [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/278435/ (дата обращения: 19.03.2022)

3. Сегментация изображения. – [Электронный ресурс]. – Режим доступа: http://mechanoid.su (дата обращения: 19.03.2022)

4. Афонасенко А.В., Обзор методов распознавания структурированных символов // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2008. – № 2 (18), часть 1. – С. 83-88.

5. Tesseract OCR. – [Электронный ресурс]. – Режим доступа: https://github.com/tesseract-ocr/tesseract (дата обращения: 22.03.2022)

6. ABBYY FineReader. – [Электронный ресурс]. – Режим доступа: https://pdf.abbyy.com/ru/finereader-pdf/ (дата обращения: 22.03.2022)

7. Cloud Vision API. – [Электронный ресурс]. – Режим доступа: https://cloud.google.com/vision/ (дата обращения: 22.03.2022)

ANALYSIS OF TEXT RECOGNITION TECHNOLOGIES FROM IMAGE

K.A. Bobrov, *Master*
V.D. Shulman, *Master*
K.P. Vlasov, *Master*
Bauman Moscow State Technical University
(Russia, Moscow)

Abstract. *The article analyzes various technologies for recognizing text from an image. Within the framework of the article, the steps of the algorithm for recognizing text from an image are considered. The main methods of the character recognition algorithm (OCR) are described, their advantages and disadvantages are given. Based on the results of the comparison, it was found that the neural network recognition method is the most promising. An analysis is made of libraries that actively use this neural network method within themselves. In conclusion, a conclusion is made about the expediency of using the considered libraries, depending on the conditions and specifics of the tasks.*

Keywords: *character recognition, segmentation, image classification, post-processing, neural network method, Tesseract technology.*