

*Илларионов А.А.,
старший оператор*

3-я научная рота Космических войск Воздушно-Космических сил

Россия, г. Красногорск

Чернов А.А.,

магистр

магистр кафедры «Робототехнические системы и мехатроника»

МГТУ им. Н.Э. Баумана

Россия, г. Москва

ОБЗОР МЕТОДОВ РАСПОЗНАВАНИЯ ТЕКСТА

***Аннотация:** в статье рассматриваются принципы построения системы оптического распознавания текста. Приводится краткое описание основных групп методов, используемых при распознавании текста, их преимущества и недостатки.*

***Ключевые слова:** оптическое распознавание символов, искусственные нейронные сети, анализ, шаблонный метод, структурный метод, признаковый метод.*

***Annotation:** The article discusses the principles of building an optical text recognition system. A brief description of the main groups of methods used in the recognition of the text, their advantages and disadvantages.*

***Key words:** optical character reader, artificial neural networks, analysis, template method, structural method, attribute method.*

***Введение.** В настоящее время для банков, промышленных и торговых предприятий все более актуальной становится проблема неэффективности бумажного делопроизводства. Для перехода к автоматизированному электронному документообороту по унифицированным и стандартизированным*

процессам, построенных на безбумажных технологиях, необходимо решить задачу распознавания текстовой информации для перевода печатных и рукописных документов в электронную форму [1].

Вышеописанная задача относится к разделу задач по распознаванию образов. У задач данного класса выделяют два основных направления:

1. Изучение и моделирование способностей к распознаванию, которыми обладают живые существа;
2. Развитие теории и методов построения устройств, предназначенных для решения отдельных прикладных задач [2].

В последние десятилетия развитие методов обработки изображений и распознавания образов для отдельных прикладных задач позволило создать системы оптического распознавания символов (OCR-системы, от англ. optical char recognition). Так, например, программа FineReader позволяет распознать печатный текст с малым количеством ошибок, требующих человеческого контроля.

Тем не менее, создание каждого нового приложения в данной области по-прежнему остается творческой задачей и требует дополнительных исследований в связи со специфическими требованиями по разрешению, быстродействию, надежности распознавания и объему памяти, которыми характеризуется каждая конкретная задача [1].

Особенности проектирования OCR-систем (Optical Character Reader).

При распознавании рукописного или печатного документа используется многоступенчатая структура, схожая с той, что использует человек для распознавания образов. В данной структуре документа выделяются семь уровней, каждый из которых содержит один и более объектов.

Пример структуры документа представлен на рисунке 1.

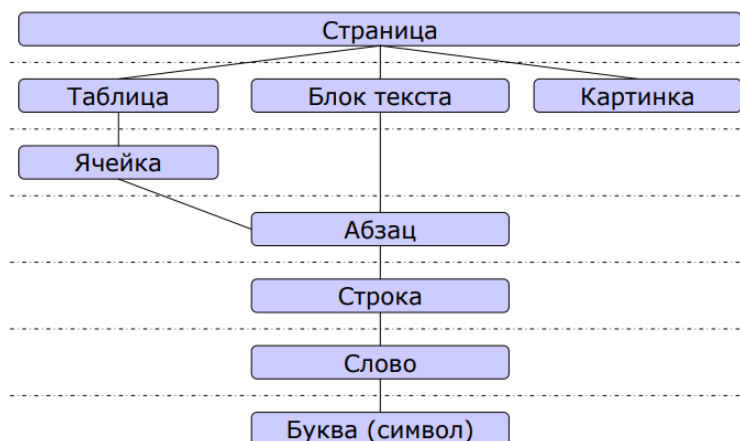


Рисунок 1. Многоуровневая структура документа

Большинство современных OCR-систем проводит анализ документа в соответствии с одним из следующих принципов: top – down («сверху вниз») или bottom – up («снизу вверх»). Например, анализ и выделение объектов в документе происходит сверху вниз, а генерация его электронной версии снизу вверх.

При распознавании рукописных и печатных символов имеется ряд проблем, связанных со следующими особенностями:

1. Разнообразие форм начертания символов;
2. Искажение изображений символов;
3. Вариации размеров и масштаба символов;
4. Изменение наклона символов;
5. Искажение формы символа за счет оцифровки изображения с «грубым» дискретом;
6. Смещением символов или частей символов относительно их ожидаемого положения в строке.

Для реализации вышеперечисленных принципов система распознавания текстов должна состоять из следующих блоков:

1. Блок сегментации элементов текста;
2. Блок предобработки изображения;
3. Блок выделения признаков;

4. Блок распознавания символов;
5. Блок постобработки результатов распознавания.

В данной статье уделяется внимание блоку распознавания символов и рассмотрению наиболее распространенных методов, которые используются при проектировании указанного блока. На данный момент выделяют три основных подхода для решения задачи распознавания: структурный, признаковый и шаблонный [3].

Шаблонные методы распознавания символов. При использовании данной группы методов отсканированное изображение переводится в растровое (поточечное), затем производится его сравнения с эталонными шаблонами, которые были сформированы в базе данных. Критерием выбора шаблона является наименьшее количество точек, отличных от исследуемого изображения. Шаблон для каждого класса получают, усредняя изображения символов обучающей выборки.

Преимуществом шаблонных методов является высокая точность распознавания дефектных символов.

Основным недостатком является зависимость от шрифта, который встречается в изображении. Используемый шрифт должен быть известен заранее, иначе будет невозможно верно распознать исследуемый символ.

Структурные методы распознавания. Данная группа методов использует информацию не о поточечном написании символа, а о его топологии: взаимном расположении отдельных составных частей символа [3, 4].

К преимуществам структурных методов распознавания можно отнести инвариантность относительно типов и размеров шрифтов [4].

Основной проблемой топологических методов является идентификация знаков, имеющих дефекты (например, разрыв линии или слияние соседних линий), а также невысокое быстродействие [3].

Признаковые методы. Методы рассматриваемой группы наиболее распространены. Анализ в них базируются на том, что изображению ставится в соответствие N – мерный вектор признаков. Распознавание заключается в

сравнении его с набором эталонных векторов той же размерности. Качество распознавания зависит от типов признаков и их количеств. Формирование вектора производится при проведении анализа подготовленного изображения, эталон для каждого класса получают аналогичным способом обработки символов из обучающей выборки.

Основными преимуществами данной группы являются простота реализации, хорошая обобщающая способность, устойчивость к изменениям формы символов, высокое быстродействие.

Из недостатков указанных методов стоит выделить неустойчивость к дефектам изображения, потеря части информации о символе на этапе извлечения признаков, отсутствия четко сформулированных правил по формированию признаков.

Альтернативные методы. На ряду с другими методами, актуальны стали методы, не требующие предварительной сегментации, такие как иерархические скрытые модели Маркова и свёрточные нейронные сети. В связи с возникновением новой волны популярности нейросетевых классификаторов они стали чаще использоваться в исследовательских работах по распознаванию текста [5]. Основным преимуществом использование нейросетевых технологий является хорошая обобщающая способность, возможность использовать контекстный анализ и распознавать символ, основываясь на окружающие его символы.

Заключение. На сегодняшний день системы по автоматическому распознаванию не достигают тех показателей, которые присуще человеческому восприятию текста, однако комбинирование различных методов улучшают показатели. Например, в OCR-системах АBBYY задействованы шесть классификаторов: растровый, признаковый, признаково-дифференциальный, контурный, структурный и структурно-дифференциальный, которые применяются в зависимости от входных параметров изображения. Точность распознавания данной системы составляет примерно 96 %.

Проанализировав существующие методы распознавания текстов, можно сделать вывод, что наиболее перспективным вариантом построения системы распознавания текста является использование сверточных нейронных сетей, которые обладают высокой обобщающей способностью, лишены недостатков вышеописанных методов, а проблема, связанная с невысокой производительностью может быть решена за счет оптимизации архитектуры сети и развития вычислительной техники.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Оптическое распознавание символов (OCR) [Электронный ресурс]. URL: [http://wiki.technicalvision.ru/index.php/Оптическое_распознавание_символов_\(OCR\)](http://wiki.technicalvision.ru/index.php/Оптическое_распознавание_символов_(OCR)) (дата обращения 28.05.19)
2. Методы распознавания текста [Электронный ресурс]. URL: <https://habr.com/ru/post/220077/> (дата обращения 28.05.19)
3. Афонасенко А.В., А.И. Елизаров, Обзор методов распознавания структурированных символов. [Электронный ресурс]. URL: <https://cyberleninka.ru/article/v/obzor-metodov-raspoznavaniya-strukturirovannyh-simvolov> (дата обращения 28.05.19)
4. Гайдуков Н.П., Савкова Е.О., Обзор методов распознавания рукописного текста. [Электронный ресурс]. URL: http://masters.donntu.org/2012/fknt/gaydukov/library/5_gaydukov.pdf (дата обращения 29.05.19)
5. Спицын В.Г., Болотова Ю.А., Фан Н.Х., Буй Т.Т.Ч., Применение вейвлет-преобразования Хаара, метода главных компонент и нейронных сетей для оптического распознавания символов на изображениях в присутствии импульсного шума. [Электронный ресурс]. URL: <https://cyberleninka.ru/article/v/raspoznavanie-simvolov-na-osnove-veyvlet-preobrazovaniya-metoda-glavnyh-komponent-i-neyronnyh-setey> (дата обращения 29.05.19).