Задача контроля качества при создании и развитии систем оптического распознавания печатного текста

Д. В. Полевой[1], О. С. Самойлов[2]

[1] *Институт системного анализа Российской академии наук,*
Россия, 117312 Москва, пр. 60-летия Октября, 9
[2] *Московский институт стали и сплавов*,
Россия, 119049 Москва, Ленинский пр., 4

Работа посвящена вопросам контроля качества работы систем распознавания. В качестве примера приведены задачи развития и использования систем оптического распознавания печатного текста.

Введение

Оптическое распознавание текста (Optical Character Recognition, OCR) — общее название для технологий и программ, преобразующих изображение текста в допускающее непосредственное редактирование электронное представление. Исходное изображение содержит печатные или написанные от руки символы. Дальнейшее рассмотрение вопросов контроля качества будет в первую очередь относиться к распознаванию печатного текста бумажных документов, но многие рассуждения и выводы применимы и для других постановок задач распознавания.

Результатом оптического распознавания является текст, который можно обрабатывать обычными способами: редактировать в соответствующих редакторах, сохранять и индексировать поисковыми системами, сравнивать с полями БД, использовать для синтеза голосовых сообщений и т. д.

Применение OCR-технологий вышло далеко за границы первоначальных задач ввода бумажных документов. Например, качественное решение задачи локализации текстовых фрагментов изображения в видеопотоке и их распознавание позволяют реализовать поиск видеозаписей по ключевым словам. Распознавание номеров автомобилей и вагонов уже используется в системах автоматического контроля и наблюдения. Существуют программные комплексы для тестирования программного обеспечения, которые опи-

раются на распознавание снимков экрана для моделирования действий пользователя при тестировании пользовательского интерфейса.

Оптическое распознавание печатного текста является промышленным инструментом, сфера применения которого все расширяется: увеличивается количество решаемых с его помощью задач, растет число пользователей. Функционал многих OCR-систем может быть использован через программный интерфейс (API), при этом существуют как свободно распространяемые системы (например, OpenOCR [1], Ocropus [2]), так и коммерческие (например, Abbyy Fine Reader [3], OmniPage [4], Readiris [5]).

Возможность выбирать из нескольких доступных OCR-систем ставит перед пользователями и прикладными программистами задачу определения лучшей или наиболее подходящей системы. Группы разработчиков технологий оптического распознавания также остро нуждаются в инструментах контроля и оценки качества OCR, особенно в ситуации совместной разработки. Таким образом, задача построения систем оценки и контроля качества работы OCR-систем является актуальной задачей.

1. Контроль качества систем распознавания печатного текста

Как любая программная компонента, система оптического распознавания имеет множество характеристик: целевая платформа функционирования, потребляемые ресурсы (память, процессорное время), устойчивость работы и т. д. Опустим анализ общих и рассмотрим специфические для задач распознавания характеристики.

Распознавание является сложным многоступенчатым процессом, в котором результаты каждого этапа существенно влияют на последующие. Создатели хорошей распознающей системы, среди прочих, решают сложную оптимизационную задачу подбора параметров алгоритмов для достижения заданного уровня качества на широком спектре документов. Распознать один–два документа и сделать выводы о качестве распознавания [6] невозможно: такой подход является лишь некоторым тестом работоспособности OCR-системы.

Основным подходом к оценке качества работы OCR-систем является сравнение текущих результатов работы с некоторыми «идеальными» (ground truth). Такие идеалы создаются в ручном или полуавтоматическом режиме и представляют собой эталонное решение задач распознавания человеком. Отклонения от эталона считаются ошибками, подсчет числа которых дает обобщенные показатели качества работы системы. Дополнительно возможны анализ типов и частоты встречаемости ошибок.

Поскольку OCR-ядро является сложной системой, его можно рассматривать как «черный» или «серый» ящик. В первом случае для оценки каче-

ства работы используют только финальные результаты распознавания. Такой способ подходит для конечного пользователя, но абсолютно не подходит для разработчиков, так как совершенно не прослеживается влияние работы отдельных подсистем на финальные результаты. Во втором случае отслеживается и отдельно оценивается качество работы каждой из компонент распознающей системы. Финальные результаты распознавания при этом являются лишь одним из контролируемых параметров. Таким образом, оценка качества для конечного пользователя является сильно упрощенным вариантом системы качества разработчика.

Для прикладного программиста лучшим вариантом оценки является набор из одной или нескольких числовых характеристик, которые можно вычислить для сравниваемых OCR-систем и из которых определить лучшую. В это же время разработчик системы распознавания должен иметь возможность детально анализировать изменения в работе отдельных алгоритмов и подсистем, получая не только количественные, но и качественные характеристики. Система контроля качества требуется как конечным пользователям OCR-библиотек, так и разработчикам последних.

2. Типовые ситуации использования оптического распознавания текста

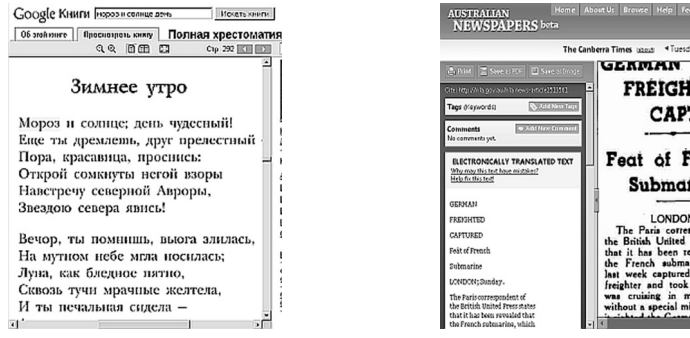
Рассмотрим основные типы задач, которые решаются с использованием оптического распознавания печатного текста.

2.1. Архивное хранение документов

В современном мире ежедневно переводится в электронную форму огромное количество бумажных документов. Количество информации в цифровой форме постоянно возрастает. Хранение электронной копии документа обладает рядом преимуществ по сравнению с хранением бумажного оригинала: сохранность внешнего вида документа на протяжении всего времени хранения, простота обмена документами, мгновенный поиск информации.

Пользователям оцифрованных документов требуются эффективные механизмы поиска, а большинство современных поисковых технологий опираются на естественное текстовое представление данных. Именно поэтому наиболее распространенной формой создания электронных копий документов является сканирование и распознавание бумажных оригиналов.

В таком варианте использования от OCR-подсистемы требуются распознанные слова и точные координаты фрагментов текста на изображении. Такая привязка текста к изображению позволит при необходимости продемонстрировать исходный фрагмент реального документа. Примерами такого подхода могут служить сервис поиска книг Google Books [7] или цифровые архивы газет [8] (рис. 1).



а) *б*)

Рис. 1. Примеры пользовательского интерфейса с совмещением результатов распознавания и исходного изображения: а — Google Books, б — Australian newspapers

В зависимости от постановки задачи основным требованием может быть либо максимальная скорость ввода, либо максимально точное воспроизведение исходного текста. В первом случае сохранение альтернатив распознавания символов и применение алгоритмов нечеткого поиска позволяют уменьшить требования к точности распознавания. Во втором варианте важна надежность автоматического детектирования ошибок распознавания, а ввод обязательно должен проходить стадию ручного контроля оператором.

2.2. Распознавание с сохранением внешнего вида документа

Очень часто требуется немного отредактировать документ, который на данный момент есть только в бумажном виде. При этом набирать вручную текст, выставлять различные текстовые параметры, отступы и т. д. — длительная и трудоемкая работа. Быстро решить поставленную задачу помогают распознающие системы, сохраняющие внешний вид документа. При этом помимо самого распознавания дополнительно определяются различные типографские параметры (размер и тип шрифта, начертание символов), сохраняются положения текста, изображений, разметка страницы, отступы и т. д. Редактирование результатов распознавания становится делом пары минут. Типичными примерами являются OCR-системы для «домашнего использования» ([3, 9]).

При таком использовании результаты распознавания всегда предоставляются пользователю «на вычитку». При достаточно хорошем качестве распознавания текста на первый план выходят возможности системы, связанные именно с анализом и восстановлением макета страницы.

2.3. Распознавание с сохранением структуры документа

Современные OCR-системы воспроизводят текст оцифрованных документов с минимальным количеством структурной информации: каждый документ является набором страниц, определены границы параграфов. Автоматическое восстановление границ глав и разделов, формирование оглавлений и ссылок внутри документов является актуальной задачей, решение которой позволит с меньшими усилиями создавать полноценные цифровые библиотеки и архивы с расширенными возможностями поиска и выборочного доступа.

Такая постановка задачи распознавания уже попадает в область «понимания документов», так как требует комплексного подхода к анализу типографической, лингвистической и семантической составляющих результатов распознавания

2.4. Ввод форм

Отдельной широкой областью применения систем распознавания является автоматизация ввода форм, т. е. документов заданной структуры и заполнения. Требования к OCR-подсистеме в таком случае сильно зависят от специфики конкретного документооборота, в рамках которого формы вводятся.

На одном полюсе находятся массовые проекты (например, по сбору маркетинговой информации), в которых важны скорость ввода и статистическая достоверность результата. При заполнении анкет люди допускают ошибки, а точность и достоверность распознавания должны сохранять общую достоверность исследования. Другим крайним случаем является ввод документов, требующих максимальной надежности, например финансовых или удостоверяющих личность. Примером такого рода систем может служить система массового ввода Cognitive Forms [10].

3. Общая схема построения инструментария контроля качества системы распознавания

Проведенный обзор показывает разнообразие областей применения распознающих систем. В разных прикладных задачах к OCR-компонентам предъявляются различные требования, поэтому более перспективным является проблемно-ориентированный подход к оценке и контролю качества распознавания. В таком подходе отправной точкой к построению критериев качества является сама прикладная задача, а не распознавание текста в отрыве от контекста его использования.

Рассмотрим схему построения инструментария контроля системы распознавания:

- анализ области применения OCR-системы и выделение существенных критериев качества;
- анализ этапов оптического распознавания и определение критических с точки зрения исходной задачи этапов;
- формирование критериев изолированной оценки работы отдельных этапов;
- создание автоматизированной системы оценки качества работы подсистем;
- создание комплексной системы контроля качества всей OCR-системы;
- изготовление инструментария и данных («идеальные» результаты).

Во время анализа области применения OCR-системы определяются границы множества распознаваемых изображений и их характерные особенности. При этом выделяются существенные параметры оценки результатов распознавания и критерии оценки системы по этим параметрам. Затем анализируется структура и основные подсистемы оцениваемой распознающей системы. При анализе следует обратить внимание на наиболее существенные с точки зрения получения оптимального ответа этапы и выделить наиболее важные подсистемы. Критерии оценки качества работы в терминах входных и выходных данных формируются на основе представлений об архитектуре системы в целом и иде-альном результате работы для выделенных подсистем.

Процедуры оценки отдельных подсистем по сформулированным критериям автоматизируются, и на их основе создается комплексный автоматизированный инструмент, который учитывает качество работы отдельных подсистем и всей системы в целом. При этом необходимо учитывать, что разные подсистемы по-разному влияют на конечный результат и к ним предъявляются разные требования, в зависимости от модели использования OCR-системы. Таким образом, для разных прикладных задач системы комплексной оценки одной и той же распознающей системы могут отличаться.

Последним и, зачастую, наиболее трудоемким этапом создания системы контроля качества является изготовление «идеалов».

4. Открытые базы изображений

Оценка качества работы OCR-системы является статистической и должна проводиться на большом количестве изображений, максимально полно описывающих множество реальных изображений в прикладной задаче. Распределение типов изображений и их характерных особенностей также должно соответствовать области применения. Изготовление вручную «идеалов» требует значительных трудозатрат, а используемые для повышения точности техники повторного ввода и сверки увеличивают стои-

Таблица 1

Описание открытых баз изображений

Название	Кол-во изображений	Разрешение (ppi)	Языки	OCR рез-ты	Формат идеала
ISRI-OCRtk	2889	200, 300, 400	Английский, испанский	Да	txt
MARG	1553	300	Английский	Да	xml
Tobaco 800	1290	150–300	Английский	Нет	xml
MediaTeam Oulu Document Database	512	300	Английский, русский, финский, немецкий	Да	binary, txt
InftyProject	—	300	Английский, французский, немецкий	Да	csv

мость еще в несколько раз. Одним из способов уменьшения их стоимости является создание повышающего производительность труда операторов инструментария. Другим вариантом является использование готовых корпусов изображений и «идеалов» (табл. 1). Проведем краткий обзор существующих в открытом доступе ресурсов.

ISRI-OCRtk — эта база является результатом пятилетнего исследования качества работы ведущих систем оптического распознавания [11]. Отсканированные в различных разрешениях и с различным качеством разнообразные бумажные документы (газеты, журналы, деловые письма, годовые отчеты и др.) сопровождаются информацией о положении и содержании текстовых блоков.

Дополнительно эта база содержит классификацию документов по типу и качеству. Особого внимания заслуживает разработанный и выложенный в открытый доступ инструментарий с подробным описанием теоретических и методологических оснований сравнительного анализа работы систем оптического распознавания текста.

MARG — эта база изображений [12] собрана в рамках проекта создания системы по оцифровке и предоставлению доступа к статьям на медицинские и биологические темы. Содержит черно-белые сканы статей на английском языке из академических биомедицинских журналов. Изображения представляют собой первые страницы статей. Идеальные результаты представлены классифицированными блоками (заголовок, список авторов,

место проведения работы, аннотация и др.) и распознанным текстом. Для всех элементов: блоков, строк, слов и символов — указаны координаты охватывающих прямоугольников.

Tobacco 800 — открытое подмножество комплекса коллекций для тестирования обработки изображений документов [13–16]. Содержит широкий спектр отсканированных на различном оборудовании с разными настройками и качеством документов. Часть изображений являются последовательными страницами коммерческих документов. «Идеалы» описывают наличие и расположение подписей и логотипов на изображениях.

MediaTeam Oulu Document Database — коллекция [17] отсканированных в цвете разнообразных, документов, изданных не позже 1978 года (списки адресов, чеки, формы, статьи, словари, карты и т. д.). Содержит информацию о физической и логической структуре всей страницы в целом (тип, номер, число колонок, язык, направление текста, число блоков, шрифт). Выделяются следующие блоки: текст (заголовок, тело, автор и др.), графика и изображение. Для каждого из блоков указывается номер, координаты, тип, язык, выравнивание и направление текста, шрифт.

Infty Project — созданная в рамках разработки специализированной системы для оптического распознавания математических текстов база [17] с искусственными изображениями формул и фрагментов текста. Общее число извлеченных символов: 662 142 из английских статей, 37 439 — из французских и 77 812 — из немецких. Описывает результаты распознавания, абсолютное и взаимное положение символов и фрагментов формул.

Заключение

Анализ практических аспектов контроля качества оптического распознавания печатного текста показывает, что построение системы контроля должно начинаться с анализа прикладных задач, решаемых с помощью OCR. В работе приведен обзор основных типовых случаев использования оптического распознавания печатного текста и предложена общая схема создания инструментов его контроля. Для оценки разнообразия документов и уменьшения стоимости разработки могут использоваться рассмотренные открытые базы изображений.

Литература

1. OpenOcr. http://www.openocr.org
2. Ocropus. http://www.ocropus.org
3. Abby FineReader. http://www.abbyy.ru
4. OmniPage. http://www.nuance.com

5. Readiris. http://www.irislink.com
6. *Acton A.* Linux OCR: A review of free optical character recognition software. http://www.eecho.info/Echo/office/linux-ocr
7. GoogleBooks. http://books.google.com/
8. Australian Newspapers Digitisation Program. http://ndpbeta.nla.gov.au/ndp/del/home
9. Cognitive Cuneiform. http://cuneiform.ru
10. *Арлазаров В. В., Постников В. В., Шоломов Д. Л.* Cognitive Forms — система массового ввода структурированных документов // Управление информационными потоками. М.: URSS, 2002. С. 35–46.
11. ISRI-OCRtk. http://www.isri.unlv.edu/ISRI/OCRtk
12. MARG. http://marg.nlm.nih.gov
13. Tobacco 800. http://www.umiacs.umd.edu/~zhugy/Tobacco800.html
14. *Lewis D., Agam G., Argamon S., Frieder O., Grossman D., Heard J.* Building a test collection for complex document information processing // In Proc. 29th Annual Int. ACM SIGIR Conference (SIGIR 2006). 2006. P. 665–666.
15. *Agam G., Argamon S., Frieder O., Grossman D., Lewis D.* The Complex Document Image Processing (CDIP) test collection project. Illinois Institute of Technology, 2006. http://ir.iit.edu/projects/CDIP.html
16. The Legacy Tobacco Document Library (LTDL). San Francisco: University of California, 2007. http://legacy.library.ucsf.edu
17. MediaTeam Oulu Document Database. http://www.mediateam.oulu.fi/downloads/MTDB
18. Infty Project Images Database. http://www.inftyproject.org/en/database.html