УДК 004.6

С.В. Кулешов, С.В. Смирнов

МЕТОДЫ СЕГМЕНТАЦИИ OCR-СИСТЕМ В ЗАДАЧАХ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ АРХИВНЫХ ДОКУМЕНТОВ

Кулешов С.В., Смирнов С.В. **Методы сегментации OCR-систем в задачах автоматической обработки архивных документов.**

Аннотация. Настоящая статья описывает сравнение современных систем оптического распознавания, проводимого с целью определить системы, наиболее точно выполняющих сегментацию документов по заранее заданным критериям; а также возможности систем по выделению различных типов областей. Анализируются результаты работы методов сегментации OCR-систем, оценивается эффективность сегментации. На основыве результатов исследования и сделанных наблюдений составлен список рекомендаций по выбору OCR-систем и методов для обработки различных типов документов.

Ключевые слова: системы оптического распознавания, методы сегментации, OCR-системы, структурный анализ документа, оптическое распознавание, оцифровка архивных документов.

Kuleshov S.V., Smirnov S.V. **Segmentation methods of OCR systems in problems of automatic processing of archival documents.**

Abstract. This paper describes the comparison of the modern optical character recognition systems aimed to find the systems, which do more precise segmentation, and to detect the capabilities of systems to allocate different types of areas. The results of the segmentation methods of OCR systems are analyzed. The effectiveness of the process of segmentation is evaluated. Based on the results of studies and observations made, recommendations to use for different types of documents are made.

Keywords: optical character recognition, segmentation, OCR systems, document layout analysis, digitization of archival documents.

1. Введение. Несмотря на широкое распространения электронного способа хранения информации, сегодня большинство информации созданной человеком, хранится либо в бумажном виде, либо на другом твердом носителе. Практически в каждой организации имеется бумажный архив, ведь даже внедрение современных систем электронного документооборота не позволяет полностью избавиться от бумажных форм хранения. Соответственно возникает ряд проблем:

— сохранения уже имеющихся документов;
— поиска;
— использования.

Типичным способом решения является создание электронного архива. Для этого организуется высокопроизводительная сеть, включающая в себя графические рабочие станции и мощные серверы ввода и обработки информации. Для ввода документов с бумажных носителей

низкого качества используются промышленные сканеры потокового ввода и соответствующие программные средства. Система обеспечивает эффективное индексирование и полнотекстовый поиск неструктурированной информации большого объема. Данные, необходимые для поиска документов, хранятся в высокопроизводительной и отказоустойчивой системе памяти, а графические образы документов — в виде изображений на носителях, характеризуемых длительным временем хранения и отказоустойчивостью.

Специфика внедрения системы электронного архивирования состоит в том, что прежде всего необходимо ввести в базу данных системы полный объем документов. Этот чрезвычайно длительный и трудоемкий процесс требует максимальной автоматизации — отстранения оператора от любого участия во вводе, распознавании, корректировке и индексировании документов

Проиллюстрируем это на конкретном примере. Допустим, бумажный архив насчитывает 50 млн. документов. На проверку одного распознанного документа, классификацию—рубрикацию, ввод атрибутов оператор среднестатистически тратит 2 минуты. Следовательно, для ввода всех документов в режиме стандартной рабочей недели потребуется 1112 лет. Кроме того, при автоматическом вводе документов основное узкое звено системы — производительность сканеров и мощность сервера, выполняющего распознавание и индексирование. С учетом оптимизации потоков подсистемы ввода можно ожидать, что аналогичный объем документов будет полностью введен за 5–15 лет.

На данный момент времени успехи по распознаванию документов впечатляют, так же как и другие значительные достижения по анализу изображений специального вида (идентификация автомобилей-нарушителей, анализ и распознавание сигналов в медицине). Однако универсальных методов обработки изображений еще не найдено, что требует проводить активную деятельность в этом направлении.

2. Роль этапа сегментации при оптическом распознавании. Автоматический или автоматизированный перевод бумажных документов в электронный вид включает в себя процесс, состоящий из двух этапов: 1) сканирования бумажных документов и 2) распознавания их содержимого с помощью специальных программ, называемых системами оптического распознавания символов (Optical Character Recognition — OCR), а также размещение полученного содержимого на устройствах хранения.

Сегментация — важный этап предобработки документа при распознавании. Его целью является разделение изображения документа на

однородные зоны, например, содержащие только текст, таблицы, графику или разделители. Этот этап весьма трудный и в общем виде не алгоритмизированный до конца для произвольных изображений. Во многих случаях точность работы OCR-системы сильно зависит от точности работы применяемых алгоритмов сегментации.

3. Цель сравнения. Целями проведения данного сравнения являются определение систем наиболее точно выполняющих сегментацию документов по заранее заданным критериям; определение возможностей систем по выделению различных типов областей; подбор наиболее подходящих типов документов для каждой системы.

В ходе сравнения необходимо определить все возможные варианты интеграции типов документов с системами с целью получить результаты структурного анализа.

Дополнительной задачей является исследование состояния развивающегося сектора систем с открытыми исходными кодами на предмет конкурентоспособности с коммерческими «гигантами».

4. Методика сравнения и критерий эффективности методов. Существует много факторов, влияющих на результаты сравнения и оценки алгоритмов и методов структурного анализа документа. В процессе подобного эксперимента в первую очередь необходимо следующее:

1) набор тестовых изображений документов;
2) эталонное описание зон сегментации для каждого изображения;
3) критерии оценки;
4) состав участников сравнения (алгоритмов или систем);
5) набор результатов сравнения результата сегментации с эталоном для каждого алгоритма по каждому изображению. Последующее суммирование результатов и принятие решений на основе полученных показателей.

В нашем случае тестовые изображения возьмем из общедоступного набора изображений в Интернете.

В роли участников будут выступать OCR-системы, обладающие функциями структурного анализа документов. Для определения претендентов рассмотрим круг из нескольких десятков современных свободно распространяемых систем, и для каждой проведем тест на первичном наборе изображений. По результатам теста отсеются системы, которые либо не выполняют сегментацию, либо не предоставляют возможность получить результаты после данного этапа, либо качество сегментации у них намного менее точное, чем у других систем. Из коммерческого сектора OCR-систем в состав участников включим

признанных лидеров — системы «ABBYY Finereader» и «Nuance OmniPage».

В качестве критериев оценки эффективности будем использовать следующие:

1) T_s — общее число всех определенных зон;
2) T_c — общее число корректно определенных зон;
3) T_{os} — общее число операций объединения, которые необходимо совершить, чтобы все сверхсегментированные зоны привести к эталону. Данный критерий отражает, насколько более сегментирован результат в сравнении с эталоном;
4) T_{us} — общее число операций разбиения, которые необходимо совершить, чтобы все объединенные зоны привести к эталону. Данный критерий отражает, насколько менее сегментирован результат в сравнении с эталоном;
5) Z_{os} — число чрезмерно разбитых зон;
6) Z_{us} — число зон, требующих разбиения на более мелкие;
7) Z_{miss} — число пропущенных зон;
8) Z_f — число ложно определенных зон;
9) Z_{ovl} — число перекрывающих друг друга зон;
10) Z_{merg} — число объединенных зон.
11) N — число неложных зон, границы которых расширены из-за объединения с шумовыми зонами. Данный критерий отражает, уровень влияния «шума» на результат работы.

Введем дополнительную меру оценки ρ — коэффициент отклонения от эталона:

$$\rho = \frac{\left| Z_{miss} \cup Z_{os} \cup Z_{merg} \right|}{|G|},$$

где $|G|$ — мощность множества G всех эталонных зон документа.

5. Участники. Рассмотрим системы, обладающие функциями сегментации и структурного анализа документа, и выбранные для участия в сравнении и анализе.

5.1. «Ocropus». Система структурного анализа документов и распознавания символов, свободно распространяемая по лицензии «Apache License, Version 2.0», с модульной организацией на основе плагинов.

В настоящее время «Ocropus» развивается под руководством ученых из Немецкого научно-исследовательского центра по искусственному интеллекту в г. Кайзерслаутерн и спонсируется корпорацией «Google». Система «Ocropus» разработана под ОС Linux.

Работа с системой возможна только через командную строку, однако при запуске в отладочном режиме промежуточные результаты отображаются через графический интерфейс.

Архитектура системы состоит из трех основных компонентов:

1) физического и логического структурного анализа документа;

2) распознавания текста;

3) статистического языкового моделирования.

Структурный анализ документа позволяет выделить в документе текстовые колонки, текстовые блоки, текстовые линии, а также порядок их чтения.

Компонент распознавания текста отвечает за распознавание каждого символа внутри текстовой строки (отметим, что строки могут быть вертикальными или с порядком чтения справа налево). Результат работы компонента — набор вариантов распознавания с вероятностными оценками, представленный в виде графа гипотез.

Языковое моделирование объединяет набор гипотез со знаниями о языках, словарях, грамматике и контексте документа, что вкупе дает наиболее достоверную оценку для выбора конечного варианта распознавания.

Остановимся более подробно на этапе структурного анализа документа. В системе реализованы следующие алгоритмы и методы:

— «RAST-Based Layout Analysis» («OcropusR»). Основной метод системы базируется на двух взаимосвязанных алгоритмах для определения пробелов и границ текста соответственно. Метод обладает следующими особенностями: работа с прямоугольными и непрямоугольными областями; малое число настраиваемых параметров; определение границ текста различной ориентации на одной странице; нечувствительность к шуму на полях документа; отсутствие глобального порога;

— «Recursive XY Cuts» («OcropusXY»). Алгоритм часто используют в сфере анализа документов. Он обладает рядом ограничений: на полях документа должен отсутствовать шум; работает только на документах с содержимым, выровненным по горизонтальным и вертикальным осям; требует обязательной установки пороговых значений. В ходе работы алгоритм формирует древовидную структуру, корнем которой является об-

рабатываемое изображение, а все листья представляют собой результат сегментации.

— «Morphological method» («OcropusM»), который можно расширить до алгоритма «segment—by—smearing», и «1 Column Projection» («Ocropus1CP») алгоритм, подходящий для анализа документов, содержащих только одну колонку текста. На примере реализации этих двух тривиальных алгоритмов удобно имплементировать собственные методы структурного анализа в системе, это алгоритмы морфологической сегментации с фиксированными размерами

Результат работы системы сохраняется в формате hOCR (HTML-OCR). Данный формат основан на языке гипертекстовой разметки с добавлением новых тэгов и атрибутов.

5.2. «OCRFeeder». Система оптического распознавания, свободно распространяемая по лицензии «GNU General Public License».

Работа с системой осуществляется через графический интерфейс или командную строку. Преимуществами графического интерфейса являются возможности предварительной обработки изображения, функции корректировки результатов структурного анализа и распознавания, возможности импорта PDF-файлов.

Система работает по упрощенному алгоритму: отсутствует этап предобработки, используются заранее предустановленные в ОС движки распознавания, в инсталляционный пакет системы включен движок «Ocrad».

Система разработана под ОС семейства Linux. Результат работы сохраняется в форматах HTML и ODT.

5.3. «Cuneiform». Система оптического распознавания, разработанная российской компанией «Cognitive Technologies». С 2008 года система распространяется свободно по лицензии «Simplified BSD License». На данный момент версии системы под ОС Windows и Linux развиваются параллельно.

Версия системы под ОС Windows представляет собой полностью законченный продукт, с графическим интерфейсом, возможностью получения изображений со сканера, функциями предобработки и корректировки промежуточных результатов, а также пакетным распознаванием.

Linux-версия системы реализована как утилита для работы только через командную строку. Главным недостатком данной версии является отсутствие поддержки определения табличных областей на этапе структурного анализа документа.

Результаты работы сохраняются в форматах hOCR, HTML, RTF, TXT.

5.4. «Finereader». Коммерческая система оптического распознавания, являющаяся одним из лидеров на рынке данной продукции. Графическая оболочка системы поддерживает только ОС Windows, для разработчиков предоставляются инструментальные средства под любые семейства ОС.

5.5. «Omnipage». Коммерческая система оптического распознавания, разработанная компанией «Nuance». По популярности и функциям ничем не уступает конкуренту «FineReader».

6. Ход выполнения и полученные результаты. Исследование проведено в области архивных документов, вследствие чего в качестве тестовых данных использован один из общедоступных наборов, схожий по своему составу с реальными изображениями документов архивных фондов. Таким образом, для проведения эксперимента выбран ряд изображений из набора «Complex document image processing (CDIP) test collection constructed by Illinois Institute of Technology». Данные изображения собраны и отсканированы с использованием различной аппаратуры. Разрешение изображений варьируется от 150 до 300 dpi, а размеры — от 1200 × 1600 до 2500 × 320 пикселей.

Результаты сравнения проверены путем экспертной оценки (табл. 1 и 2).

В результатах работы системы «Cuneiform» содержится наибольшее число пропущенных зон, которые делятся в равной степени на полностью и частично неопределенные. Характерным показателем также является существенное число зон, требующих разбиения, что связано в большинстве случаев с объединением сплошных полос шума и значимых областей. Данная система является единственной, в результатах работы которой встретились случаи выделения текстовых областей как графических.

Система «Finereader» наименее всех чувствительна к различным проявлениям шумовых эффектов, в связи с этим имеет самый минимальный показатель ошибочного выделения областей. Чрезмерное объединение областей привело к наименьшему значению показателя общего числа выделенных областей и значительному числу недостаточно сегментированных зон. Внушительное значение коэффициента отклонения от эталона является следствием того же чрезмерно крупного объединения значимых зон в одну, однако это не влияет на последующую логику работы алгоритмов распознавания системы.

Таблица 1. **Результаты работы систем**

OCR	T_s	T_c	T_{os}	T_{us}	Z_{os}	Z_{us}	Z_{merg}	Z_{miss}	Z_f	Z_{ovl}	N
Cuneiform	66	31	9	45	8	21	51	5	4	0	4
Finereader	43	17	6	62	3	14	78	1	3	6	0
OCRfeeder	51	13	2	76	2	15	82	0	22	0	3
Ocropus1CP	98	11	54	57	14	23	58	1	5	0	28
OcropusM	413	46	113	27	29	19	22	0	224	2	3
OcropusR	183	52	92	11	27	10	17	2	5	7	2
OcropusR1	131	15	80	60	18	27	57	2	12	22	6
OcropusXY	133	39	18	51	8	15	55	0	54	0	10
Omnipage	90	50	5	37	5	6	41	2	3	0	1

Примечание: все показатели нормализованы по общему числу эталонных зон и представлены в процентах.

Таблица 2. **Коэффициенты отклонения от эталона.**

OCR	**ρ**
Cuneiform	0,59
Finereader	0,77
OCRfeeder	0,89
Ocropus1CP	0,57
OcropusM	0,53
OcropusR	0,51
OcropusR1	0,65
OcropusXY	0,65
Omnipage	0,65

В системе «OCRFeeder» при наличии сплошных шумовых полос на полях документа алгоритм работы сбивается, и вся область изображения принимается за одну. Даже в случае отсутствия шумовых дефектов система стремится объединить все области в одно целое. Как следствие эта система обладает максимальными значениями показателя наличия объединенных зон и коэффициента отклонения от эталона, а также максимальным числом требуемых операций разбиения.

В системе «Ocropus» метод «1CP» является «одноколоночным» и чувствителен к проявлениям шумовых дефектов, в связи с этим в результатах работы встречается наибольшее число областей, границы которых расширены до объединения с зонами шума. То есть все пространство между крайними граничными областями одной горизонтальной проекции рассматривается как одно целое независимо от того, значимые это области или нет.

Метод «OcropusM» лидирует по суммарному числу выделенных и по числу ошибочно определенных областей. Данное лидерство обусловлено выделением каждого шумового дефекта как самостоятельной области. Максимальное значение показателя излишне сегментированных зон объясняется посимвольным разбиением вертикально ориентированного текста.

В методе «RAST» системы «Ocropus» особенностью является детальное сегментирование областей — все текстовые строки внутри одной области определяются как самостоятельные сегменты. В дополнение, данный алгоритм в отличие от всех других рассмотренных свободно распространяемых систем практически не чувствителен к шуму. Две данные особенности обусловливают максимальное значение показателя корректно определенных областей, минимальное число чрезмерно объединенных областей и как следствие минимальное значение коэффициента отклонения от эталона.

Метод «RAST1» системы «Ocropus» в целом схож по результатам работы с методом «RAST». Различные значения показателей между данными методами объясняются тем, что метод «RAST1» является «одноколоночным».

Метод «Recursive XY Cuts» системы «Ocropus» выделяется среди всех других методов данной системы крупным и «грубым» сегментированием, встречаются области, разделенные и объединенные в горизонтальных и вертикальных проекциях одновременно.

Система «Omnipage» нечувствительна к шумовым дефектам и, в отличие от своего конкурента — системы «Finereader», — более детально выделяет области, что в итоге привело к большому числу корректно определенных сегментов.

7. Рекомендации. Основываясь на результатах исследования и наблюдениях, можно составить список рекомендаций по выбору OCR-систем и методов для различных типов документов (табл. 3):

- для документов с отсутствием шума и небольшим углом наклона можно использовать методы «Recursive XY Cuts»« и «Morphological method» как наиболее быстрые и простые в реализации;
- документы, состоящие из одной колонки текста и не содержащие шума, можно обработать методами «1 Column Projection» и «RAST1-Based Layout Analysis»;
- метод «RAST-Based Layout Analysis» будет хорошим выбором для обработки документов, состоящих из горизонтально расположенных текстовых областей и таблиц;

— вертикально ориентированные текстовые области наиболее точно определяются системами «Cuneiform», «Finereader», «OcropusXY» и «Omnipage»;
— Системы «Finereader» и «Omnipage» обрабатывают все типы документов, но при этом система «Omnipage» точнее определяет табличные области.

Таблица 3. **Типы областей, поддерживаемые системами**

OCR	Горизонтальный текст	Вертикальный текст	Таблица	Графика
Cuneiform	+	+	–	+
Finereader	+	+	+	+
OCRfeeder	+	+/–	–	–
Ocropus1CP	+	+/–	–	–
OcropusM	+	+/–	–	+/–
OcropusR	+	–	+/–	–
OcropusR1	+	–	–	–
OcropusXY	+	+	–	+/–
Omnipage	+	+	+	+

Примечание: «+» — область выделяется корректно; «–» — область выделяется некорректно; «+/–» — область выделяется, но не единым блоком.

8. Заключение. В результате проведения исследования, направленного на определение систем наиболее точно выполняющих сегментацию документов по заранее заданным критериям был определен набор методов: «1 Column Projection», «Morphological method», «RAST—Based Layout Analysis», «RAST1—Based Layout Analysis», «Recursive XY Cuts».

В зависимости от области использования программного продукта различные критерии будут применяться для выбора подходящего метода или системы. Используя эти критерии, мы проанализировали сильные и слабые стороны девяти участников сравнения.

Исследование показало, что методы «Recursive XY Cuts», «Morphological method», «1 Column Projection» некорректно работают на документах с большим количеством шума, для «одноколоночных» документов лучше всего подходит метод «RAST1–Based Layout Analysis», а метод «RAST–Based Layout Analysis» занимает безоговорочное лидерство при обработке текстовых документов.

Среди свободно распространяемых систем по числу типов обрабатываемых документов лидирует система «Cuneiform», а система

«OCRfeeder» продемонстрировала самые непредсказуемые варианты сегментации документов.

На данный момент коммерческие системы, несомненно, заслуживают наивысшую оценку, хотя по возможностям настройки, адаптации и интеграции уступают развивающимся системам с открытыми исходными кодами.

Литература

1. *Antonacopoulos A., Bridson D.* Performance Analysis Framework for Layout Analysis Methods // Proc. of the 9th Intern. Conf. on Document Analysis and Recognition (ICDAR2007). Curitiba, Brazil, September 2007. С. 1258–1262.
2. *Chaudhuri B. B.* Digital Document Processing: Major Directions and Recent Advances. L.: Springer, 2007.
3. *Berkner K., Likforman-Sulem L.* Special issue on document recognition and retrieval 2009 // Intern. J. on Document Analysis and Recognition. 2010. № 2. С. 77–78.

Кулешов Сергей Викторович — канд. техн. наук, старший научный сотрудник лаборатории автоматизации научных исследований Учреждения Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН). Область научных интересов: инфологические информационные системы, аналитический мониторинг Интернет, обработка видео данных. Число научных публикаций — 60. kuleshov@iias.spb.su, www.sial.iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)323-5139, факс +7(812)328-4450.

Kuleshov Sergey Viktorovich — Ph.D. in Technics, senior researcher, Laboratory Automation Research, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: infological information systems, analytical monitoring of the Internet, processing of video data. The number of publications — 60. kuleshov@iias.spb.su, www.sial.iias.spb.su; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)3235139, fax +7(812)3284450.

Смирнов Сергей Владимирович — соискатель ученой степени канд. техн. наук; Учреждение Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН (СПИИРАН). Область научных интересов: разработка системы семантическая предобработки изображений документов. Число научных публикаций — 1. serge.smir@gmail.com; СПИИРАН, 14-я линия В.О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7 911 2430840. Научный руководитель — С.В. Кулешов.

Smirnov Sergey Vladimirovich — competitor of Ph.D. in Technics; St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: development of a system of semantic preprocessing of document images. The number of publications — 1. serge.smir@gmail.com; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7 911 2430840. The scientific adviser — S.V. Kuleshov.

Рекомендовано лабораторией автоматизации научных исследований СПИИРАН, зав. лаб., д-р техн. наук., проф. В.В. Александров .
Статья поступила в редакцию 23.01.2011.

РЕФЕРАТ

Кулешов С.В., Смирнов С.В. **Методы сегментации OCR-систем в задачах автоматической обработки архивных документов.**

В статье раскрывается актуальная тема автоматизации процессов ввода, распознавания и индексирования документов при комплектовании электронного архива. Основной акцент делается на важность этапа сегментации при оптическом распознавании документов. Ведь во многих случаях, точность работы систем оптического распознавания (OCR) сильно зависит от точности работы применяемых алгоритмов сегментации.

Путем проведения анализа и сравнения, существующих OCR-систем выделяется набор используемых методов сегментации. Целями сравнения являются определение систем, наиболее точно выполняющих сегментацию документов по заранее заданным критериям; определение возможностей систем по выделению различных типов областей; подбор наиболее подходящих типов документов для каждой системы.

Для оценки алгоритмов и методов структурного анализа документов применялись такие критерии, как число определенных зон, корректно определенных зон, чрезмерно разбитых зон, пропущенных зон и др. Также вводится дополнительная мера оценки — коэффициент отклонения от эталона.

В качестве участников исследования выбраны системы с открытыми исходными кодами «Ocropus», «OCRFeeder» и «Cuneiform», а также коммерческие системы «Finereader» и «Omnipage».

Для проведения эксперимента использован ряд изображений из набора «Complex document image processing (CDIP) test collection constructed by Illinois Institute of Technology».

По результатам проведенного исследования определен набор методов сегментации: «1 Column Projection», «Morphological method», «RAST—Based Layout Analysis», «RAST1—Based Layout Analysis», «Recursive XY Cuts».

Выявлено, что методы «Recursive XY Cuts», «Morphological method», «1 Column Projection» некорректно работают на документах с большим количеством шума, для «одноколоночных» документов лучше всего подходит метод «RAST1—Based Layout Analysis», а метод «RAST—Based Layout Analysis» занимает безоговорочное лидерство при обработке текстовых документов.

Среди свободно распространяемых систем по числу типов обрабатываемых документов лидирует система «Cuneiform», а система «OCRfeeder» демонстрирует самые непредсказуемые варианты сегментации документов.

Таким образом, на данный момент коммерческие системы, несомненно, заслуживают наивысшую оценку, но по возможностям настройки, адаптации и интеграции уступают развивающимся системам с открытыми исходными кодами.

SUMMARY

Kuleshov S.V., Smirnov S.V. **Segmentation methods of OCR systems in problems of automatic processing of archival documents.**

The article explains the topic of the automation of input, recognition and indexing of documents in manning the electronic archive. Emphasis is placed on the importance of the segmentation stage in the process of OCR documents. Indeed, in many cases, the accuracy of the system optical character recognition (OCR) strongly depends on the accuracy of the algorithms used for segmentation.

Through analysis and comparison of existing OCR systems is allocated a set of methods used for segmentation. The objectives of the comparison is to identify the systems most accurately performing segmentation of documents to a predefined criteria, the definition of system capabilities to allocate different types of areas, selecting the most suitable types of documents for each system.

To evaluate the algorithms and methods of structural analysis of documents used criteria such as the number of certain zones, the number of well-defined zones, the number of excessively broken zones, the number of missed areas and others. Also, an additional measure estimates c — coefficient of deviation from the standard is introduced.

Participants in the comparison are system of open-source: «Ocropus», «OCRFeeder», «Cuneiform», and commercial systems: «Finereader», «Omnipage".

For the experiment used a number of images from the set of «Complex document image processing (CDIP) test collection constructed by Illinois Institute of Technology».

The research identifies a range of methods of segmentation: «1 Column Projection», «Morphological method», «RAST-Based Layout Analysis», «RAST1-Based Layout Analysis», «Recursive XY Cuts".

The report states that the method «Recursive XY Cuts», «Morphological method», «1 Column Projection» improperly working on documents with a lot of noise, for «single-column» document best method «RAST1-Based Layout Analysis», and the method of «RAST-Based Layout Analysis» is the unconditional leader in the processing of text documents.

Among the freely distributed systems for many types of documents processed by the lead system «Cuneiform», and the system «OCRfeeder» demonstrates the unpredictable variations segmentation of documents.

Thus, at present, commercial systems, certainly deserve the highest rating, but the customization, adaptation and integration of inferior developing open source systems.