СРАВНЕНИЕ OCR-СИСТЕМ НА ОСНОВЕ ТОЧНОСТИ АНАЛИЗА ИЗОБРАЖЕНИЯ

А.И. Андрианов,
бакалавр кафедры «Распознавание изображений и обработка текста»,
Московский Физико-Технический Институт (Государственный Университет).
Адрес: 129301, Россия, Москва, а/я 49, компания ABBYY ,
e-mail: atwice@yandex.ru.

В статье описаны методы оценки и сравнения OCR-систем по качеству анализа графического изображения. Приведена модель разметки структуры графического изображения, содержащего текст. Предложены два вида сравнительнойоценки блока анализа OCR-системы. На основании предложенных оценок проведено сравнение двух OCR-систем, разработанных компанией ABBYY: FineReader 8 и FineReader 9.

Ключевые слова: OCR, анализ изображения, распознавание текста, оценка и сравнение OCR-систем.

Введение

В современном деловом документообороте широкое распространение получили безбумажные информационные технологии и системы электронного документооборота. Основной проблемой при переходе на новые технологии является перевод информации с бумажных носителей в электронную форму. Эта задача решается системами оптического распознавания символов (OCR-системы, от англ. optical char recognition). Большинство OCR-систем работают с растровым изображением, которое получено через факс-модем, сканер, цифровую фотокамеру или другое устройство. Результатом работы системы распознавания текста является отформатированный документ, сохраненный на компьютере в одном из распространенных форматов данных.

В связи с тем, что на рынке представлено достаточно много систем OCR, возникает задача сравнения и оценки данных систем. Основным критерием оценки систем оптического распознавания символов традиционно является точность распознавания. Однако на сегодняшний день точность практически всех систем распознавания текста превышает 99,9%. Эта точность фактически означает, что на каждую страницу печатного текста приходится в среднем 1-2 неверно распознанных символа. Следовательно, результат оцифровки любой OCR-системой пока требует человеческого контроля. Таким образом, точность распознавания является недостаточным критерием для сравнения OCR-систем.

Важной подзадачей оцифровки электронного документа является *анализ* изображения. На этапе *анализа* в графическом документе OCR-системы выделяет зоны разнотипной информации и сохраняет расположение и размеры этих областей. Текстовые области анализируются дополнительно. При этом выделяются отдельные строки текста. Несмотря на кажущуюся простоту, это не такая очевидная задача, так как на практике неизбежны перекос изображения страницы или фрагментов страницы при сгибах. Даже небольшой наклон приводит к тому, что левый край одной строки стано-

вится ниже правого края следующей, особенно при маленьком межстрочном интервале. В результате возникает проблема определения строки, к которой относится тот или иной фрагмент изображения. Например, для букв j, й, ё при небольшом наклоне уже сложно определить, к какой строке относится верхняя (отдельная) часть символа (в некоторых случаях ее можно принять за запятую или точку).

На этом же этапе анализируется структура таблиц. После анализа распознаванию будут подлежать конкретные строки, а также области ячеек в таблицах. Неклассифицированные области, содержащие графическую информацию, система распознавания помечает как «изображение». Изображение не подлежит распознаванию, а переносится в целевой документ с сохранением масштаба и положения в документе.

От результатов анализа изображения так же зависит другой этап оцифровки – *синтез*. На этапе *синтеза* по атрибутам символов система OCR восстанавливает шрифт текста, в частности начертание (полужирный, курсив), размер, цвет. Текст форматируется в соответствии с расположением областей, полученным во время *анализа*. То есть создается разметка для колонок, задаются отступы строк, межстрочные интервалы и т.п.

Таким образом, от результата *анализа* зависят два других этапа оцифровки документов: распознавание и синтез. В данная статья рассмотрен метод оценки качества анализа графического изображения, основанный на типичных ошибках анализа.

Анализ графического документа

Под задачей анализа понимают следующую задачу.

На входе дан графический документ. На выходе имеется разметка документа. В общем случае, документ имеет произвольную форму. Документ может содержать информацию разного рода: отдельные символы, текст, форматированный текст, картинки, таблицы, диаграммы, штрих-код и графики.

Результатом задачи анализа должна быть разметка документа. Разметка представляет собой набор областей того или иного типа. Области текста, штрих-кода или картинки характеризуются только типом и границами. Область таблицы должна также содержать разметку для ячеек.

Пример.

Для текстового документа (*рисунок 1*) должна быть получена разметка (*рисунок 2*).

Ошибки в задаче анализа

Примером неправильного анализа документа могут служить следующие ошибки:

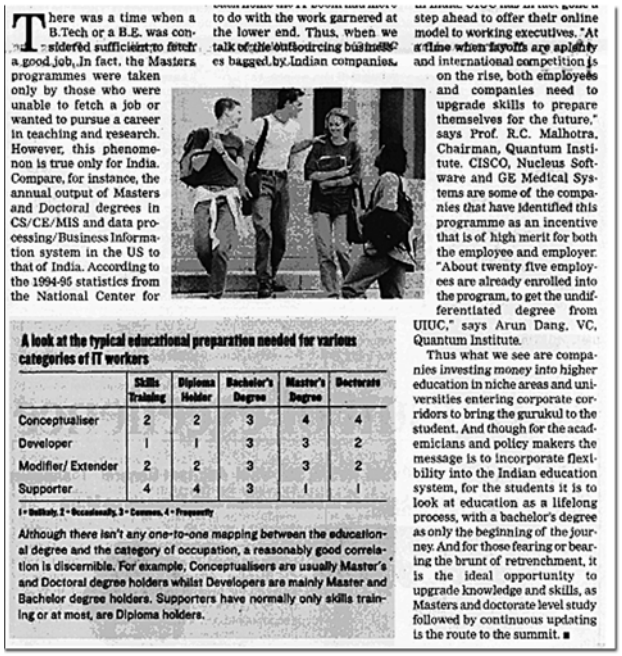
✦ Маркер маркированного списка в строке потерян. Теряется информация о том, что данный текст является списком.

✦ Потеря двоеточия в конце строки.

✦ Подписи в диаграмме ошибочно принимаются за часть картинки.

✦ В следующем примере цвет букв сливается с внешним фоном и строка целиком теряется.

EXAMPLE



There was a time when a B.Tech or a B.E. was considered sufficient to fetch a good job. In fact, the Masters programmes were taken only by those who were unable to fetch a job or wanted to pursue a career in teaching and research. However, this phenomenon is true only for India. Compare, for instance, the annual output of Masters and Doctoral degrees in CS/CE/MIS and data processing/Business Information system in the US to that of India. According to the 1994-95 statistics from the National Center for

to do with the work garnered at the lower end. Thus, when we talk of the outsourcing business bagged by Indian companies,

and international competition is on the rise, both employees and companies need to upgrade skills to prepare themselves for the future," says Prof. R.C. Malhotra, Chairman, Quantum Institute. CISCO, Nucleus Software and GE Medical Systems are some of the companies that have identified this programme as an incentive that is of high merit for both the employee and employer. "About twenty five employees are already enrolled into the program, to get the undifferentiated degree from UIUC," says Arun Dang, VC, Quantum Institute.

step ahead to offer their online model to working executives. "At a time when layoffs are aplenty

Thus what we see are companies investing money into higher education in niche areas and universities entering corporate corridors to bring the gurukul to the student. And though for the academicians and policy makers the message is to incorporate flexibility into the Indian education system, for the students it is to look at education as a lifelong process, with a bachelor's degree as only the beginning of the journey. And for those fearing or bearing the brunt of retrenchment, it is the ideal opportunity to upgrade knowledge and skills, as Masters and doctorate level study followed by continuous updating is the route to the summit. ■

A look at the typical educational preparation needed for various categories of IT workers

	Skills Training	Diploma Holder	Bachelor's Degree	Master's Degree	Doctorate
Conceptualiser	2	2	3	4	4
Developer	1	1	3	3	2
Modifier/ Extender	2	2	3	3	2
Supporter	4	4	3	1	1

1 = Unlikely, 2 = Occasionally, 3 = Common, 4 = Frequently

Although there isn't any one-to-one mapping between the educational degree and the category of occupation, a reasonably good correlation is discernible. For example, Conceptualisers are usually Master's and Doctoral degree holders whilst Developers are mainly Master and Bachelor degree holders. Supporters have normally only skills training or at most, are Diploma holders.

Рис. 1

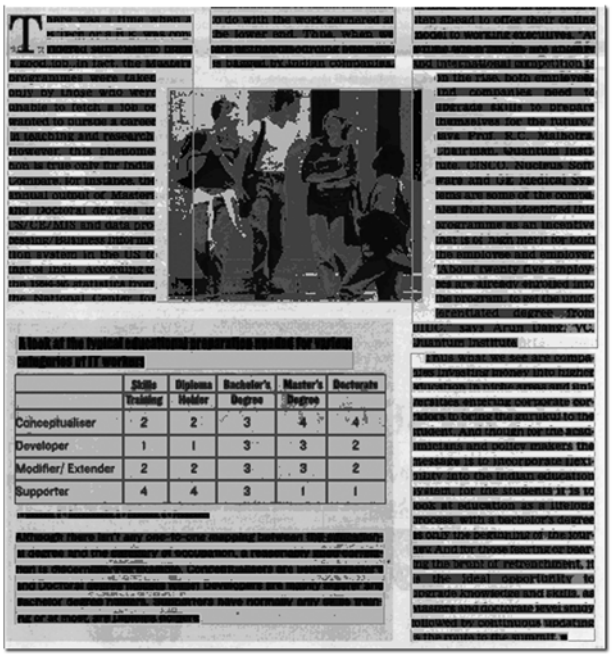


Рис. 2

✦ Несколько колонок в тексте ошибочно объединены в одну.

Существует много разных типов ошибок. При этом разные типы ошибок неравнозначны. Одни ошибки приводят к совершенно неверному распознаванию текста, другие же просто приводят к незначительным потерям информации. Ошибки также неравнозначны по времени, затрачиваемому на их исправление. Например, маркированный список можно создать уже после распознавания текста за несколько секунд, однако неправильно выделенную картинку сложной формы придется выделять вручную заново, при этом необходимо заново распознавать страницу.

Для того, чтобы можно было сравнивать результаты анализа графического документа необходимо во-первых, иметь эталонную разметку документа, а во-вторых, классифицировать ошибки. Обе эти задачи решает разметка структуры документа, описанная ниже.

Рис. 3 Разметка структуры документа

Разметка структуры служит для того, чтобы абстрагироваться от конкретных реализаций выделения зон документа. Например, в ABBYY FineReader зона текста выделяется целым текстовым блоком с горизонтальными и вертикальными границами, при этом пользователь не видит границы выделения конкретных строк (*рис. 3)*. Для оценки качества анализа такой формат разметки плох.

На *рис. 3* области 1, 2, 3 – текстовые, 4 – область картинки. Видно, что часть картинки (облако) попадает в текстовую область 3, однако это не приведет к ошибке, т.к. строки символов будут распознаны как текст, а облако не попадет в эти области. Тем не менее, при оценке мы точно должны знать, когда система OCR пытается распознать картинку, так как это является ошибкой.

Другой пример – необходимость сравнивать качество анализа документа систем разных производителей. Возможно, форматы разметки таких систем будут отличаться, тогда просто необходимо привести их к одному виду.

Наконец, в любом случае, правильная разметка не уникальна. Захват дополнительной пустой области в любую другую не является ошибкой. При выделении колонок также неважно, в каком месте пройдет разделение между колонками. Если колонки распознаны отдельно – ошибки нет.

Примитивы модели.

В предложенной модели разметки графического документа всего 6 видов блоков разметки:

2. Линия текста;
3. Картинка;
4. Таблица;
5. Штрих-код;
6. Разделитель;
7. Мусор.

Каждый блок разметки, кроме вида «таблица», представляет собой прямоугольник. Внутри одного такого прямоугольника содержится только информация данного типа.

Линия текста – это блок, в который заключена одна строка текста, содержащая единый смысл. Например, в тексте, разделенном на 2 колонки линия текста – это одна строка из одной колонки.

Картинка – блок, содержащий графическую информацию, которая должна быть перенесена в электронный документ, но не подлежащая распознаванию. Часто картинки имеют не прямоугольные очертания. В таком случае, область картинки должна быть помечена несколькими, возможно, пересекающимися элементами разметки вида «картинка».

Таблица – в блоке этого вида разметки указываются не только границы таблицы, но и внутренняя структура, т.е. указываются все столбцы и строки таблицы, а также объединенные ячейки (если есть).

Штрих-код – блок, изображение внутри которого должно быть распознано, как штрих код. С точки зрения дальнейшего распознавания текста, штрих-код – это просто картинка, однако если система умеет распознавать штрих-код, то необходимо оценивать качество выделения данной зоны.

Разделитель – это служебный элемент, обычно в виде тонкой линии. Смысл его в том, что один текстовый блок не может содержать строки одновременно справа и слева от разделителя. Проще говоря, строки текста не должны пересекать разделитель.

Мусор. Графический документ может содержать элементы, не несущие информацию. Шумы и мусор нужно выделять специальным блоком, чтобы эта информация не подлежала распознаванию и не переносилась в распознанный документ.

Классификация ошибок на основе унифицированной разметки

На основе примитивов модели выделяются следующие ошибки, характерные системам OCR при анализе документа:

Незаконченная строка. Возникает, если анализатор не целиком выделил строку текста, то есть часть строки потеряна. Примером такой ошибки является неучтенный при анализе маркер маркированного списка. Маркер является частью строки, однако анализатор часто теряет его. Подобная ошибка происходит со знаками препинания в конце строки. Другой пример этой ошибки – знак подчеркивания «земля», как поле для заполнения. Обычно текст находится по обе стороны от подчеркивания на одной строке, но анализатор, подлежащий оценке, считает такую конструкцию разными блоками текста.

Картинка вместо текста. Возникает, если разметкой выделен текст, а испытуемый анализатор считает эту строку частью картинки. Такие ошибки часто возникают, если фон картинки и текста одного цвета, или текст находится очень близко к картинке, например подпись к диаграмме или фотографии.

Неправильное выделение абзаца. Обычно возникает сразу несколько ошибок данного типа, когда анализатор объединяет несколько абзацев. Чаще всего появляется, если границы текста имеют смещения, например, из-за картинки внутри текста.

Потеря строки. Возникает, если анализатор не нашел целую линию текса. Такая ошибка часто возникает, если буквы в колонтитулах очень маленькие или ошибочно выделена таблица.

Пересечение разделителя. Возникает, если какой-либо блок разметки пересекает разделитель. Эта ошибка признак того, что текст из разных колонок объединен в один блок.

Текст вместо картинки. Возникает, если анализатор считает текстом часть области картинки. Ошибка часто появляется, если в картинке присутствуют буквы или при анализе диаграмм.

Захват региона штрих-код. Возникает если часть области «штрих-код» помечена как текстовая область.

Захват мусора. Возникает если часть области «мусор» помечена как текстовая область.

Ошибка поиска таблицы. Возникает, если анализатор пропускает таблицу или «находит» несуществующую таблицу. Чаще всего появляется, если в картинке содержатся клетки или текст организован в виде похожем на таблицу.

Ошибка анализа таблицы. Возникает, если в таблице неправильно выделены границы ячеек. Чаще всего появляется, если таблица не регулярная (часть ячеек объединены или каждая ячейка содержит разнородную информацию, например, разными шрифтами).

Разметка структуры документа является средством оценки модуля анализа OCR-систем. Чтобы оценить качество анализа, система OCR при распознавании, основываясь на результатах анализа, генерирует унифицированную разметку, соответствующую результатам своего анализа. Затем эта разметка сравнивается с разметкой введенной вручную. На большом пакете документов подсчитывается количество ошибок разного вида. На основе информации об этих ошибках составляется оценка анализатора системы OCR.

Положительной чертой данной разметки является то, что она сглаживает возможные неоднозначности разметки. Очевидно, что не существует единственно правильной разметки. Если система OCR захватит в текстовый блок пустое пространство, то это не повлечет ошибок. Такая ситуация предусматривается унифицированной разметкой так: вручную выделяются только строки текста, если эти строки целиком попали в блок, который выделила система OCR, ошибки нет, однако если часть строки не попала в блок, значит она не будет распознана – это ошибка.

Другой пример – две колонки. Вручную между колонками должен быть вставлен блок «Разделитель». Если система OCR выделила блок, пересекающий разделитель, это значит, что она не распознала в тексте 2 колонки. В противном случае, неважно как границы разных колонок расположены между собой. Колонки распознаны отдельно, ошибки нет.

Также к положительным чертам предложенной модели следует отнести возможность автоматической генерации такой разметки. Действительно, любая система распознавания выделяет строки текста, абзацы, таблицы и картинки. Также любая область может быть представлена в виде объединения прямоугольников.

Однако у данной модели разметки есть также отрицательные стороны. Во-первых, все блоки представляют собой прямоугольник. Это играет важную роль при составлении разметки для картинок. Очень многие картинки в современных печатных изданиях имеют диагональные или фигурные границы. Если текст расположен непосредственно вблизи картинки, то выделение такой области может занять у оператора большое количество времени. Однако, вертикальные и горизонтальные границы вполне объяснимы. С такой разметкой значительно проще вычислить, входит ли одна об-

ласть в другую. Иначе, вычислительная сложность алгоритмов проверки повышается.

Во-вторых, часто приходится сталкиваться с логотипами. Логотипы с одной стороны – это картинки, с другой стороны в них может присутствовать текст, который системы OCR может распознать. Если пометить логотип блоком картинка, а в нем присутствует текст, то возникает ошибка «Захват картинки». Если же пометить весь логотип, как текстовое поле, то будет потеряна картинка и, соответственно, суть логотипа. Однако даже если такая ошибка будет встречаться при распознавании довольно часто, она может быть легко исправлена в системе OCR выделением логотипа как картинки. Таким образом, можно считать этот минус несущественным.

Инструмент для задания разметки структуры документа

Компанией ABBYY был специально разработан инструмент для задания описанной разметки. Он называется BatchAnalyzer. Эта программа позволяет не только интерактивно задавать разметку для целого пакета документов, но также вызывать блок анализа одной из версий системы OCR FineReader 7, 8 или 9. После анализа, BatchAnalyzer автоматически сверяет разметку полученную системой OCR и заданную вручную оператором, считает количество ошибок каждого вида и даже может выделить ошибочно распознанные области. В программе BatchAnalyzer также предусмотрена возможность сравнения двух результатов анализа, сопоставления ошибок и просмотр отличий результатов.

С помощью инструмента BatchAnalyzer было проведено сравнение работы модуля анализа двух OCR-систем: FineReader 8 и FineReader 9.

Пакет данных для сравнения OCR-систем

К пакету документов, на котором проводится сравнение OCR-систем предъявляется требования содержать достаточно сложнее документы, чтобы были возможны ошибки анализа всех типов. Требовалось выявить как можно больше промахов различных систем OCR. Однако в пакете не должны содержаться документы нестандартной структуры, такие как описанные ниже в разделе «Одна строка – одна ошибка» (*см. рис. 5*).

Пакет содержал 300 графических документов из различных источников. Большинство документов - отсканированные страницы из журналов и книг. Также пакет содержал снимки экранов интернет страниц. Поскольку часть документов была экономической тематики, а другая часть содержала информацию о компьютерах, то неизбежно в документах присутствовали графики, диаграммы, а также изображения со снимками окон компьютерных программ. В пакете также присутствовали таблицы разной сложности (с сеткой и без), были таблицы со слиянием ячеек. Присутствовал текст, оформленный в виде таблицы, но, по сути, таблицей не являющийся, например содержание книги. В журнальных статьях присутствовали картинки с непрямоугольными краями.

Итак, графические документы были довольно сложны для анализа. FineReader 9 на всем пакете делал ошибки всех вышеперечисленных типов. Таким образом, можно считать, что пакет документов удовлетворяет требованиям.

Целевой метод сравнения результатов анализа

Целевой метод предполагает выбор системы, совершающей наименьшее количество ошибок заданного типа. Пользователь, который имеет дело с оцифровкой большого количества документов, содержащих таблицы, не заинтересован в OCR-системе, которая производит анализ лучше других систем. Такому пользователю нужна OCR-система, которая лучше других систем умеет анализировать именно табличные данные. Пользователю, который оцифровывает журнальные статьи, нужна OCR-система, которая не теряет текст около картинки. Таким образом, получаем оценку, отвечающую нуждам пользователя, работающего с однотипными данными.

Таблица 1.

Количество ошибок разных типов в сравниваемых версиях FineReader.

Тип ошибки	Количество ошибок	
	FineReader 8	FineReader 9
Незаконченная строка	885	592
Картинка вместо текста	0	284
Неправильный абзац	327	145
Потеря строки	410	1556
Пересечение разделителя	35	28
Текст вместо картинки	745	440
Лишний текст	66	45
Потеря таблицы	94	25
Лишняя таблица	0	35

В *табл. 1* приведено количество ошибок в пакете из 300 документов для систем FineReader 8 и FineReader 9.

Как описывалось выше, суммарное количество не обусловлено случайными «выпадами» количеств ошибок того или иного анализатора. Количество ошибок каждого типа растет в среднем пропорционально размеру пробного пакета. Таким образом, данные из таблицы действительно отражают реальную картину распределения ошибок по типам.

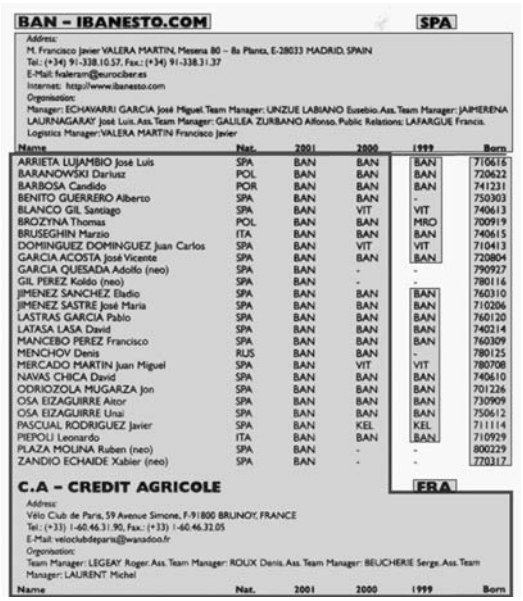
Сравним общее количество ошибок анализа по каждому из типов. Из результатов видно, что анализатор FineReader9 делает больше ошибок типа «Потеря строки», однако другие ошибки в результате работы встречаются реже.

Нули в строках «Картинка вместо текста» и «Лишняя таблица» не говорят о том, что анализатор FineReader8 правильно находит картинки и не делает лишних таблиц. Этих ошибок нет, потому что этот анализатор просто не выделяет таблицы и картинки. Становится понятно, почему нужно учитывать не только количество сделанных ошибок, но и количество правильно найденных элементов данного типа.

Метод «Одна строка – одна ошибка»

На первый взгляд может показаться, что объективной интегральной оценкой качества анализа графического изображения могла бы выступать сумма или взвешенная сумма ошибок, допущенных модулем анализа OCR-системы. Однако, это не так. Ошибки, основанные на модели разметки структуры документа не аддитивны. Например, если документ состоит из одной таблицы на целый лист, то разметкой такого документа является один блок «таблица». Соответственно, если анализатор ошибается и не находит таблиц, то оценка учитывает эту ошибку, как одну ошибку типа «Поиск таблицы». Однако таблица состояла из сотни ячеек, в каждой по строке текста. Некоторые строки в дальнейшем не будут распознаны, между другими будут потеряны связи, возможно в один блок будут объединены строки из разных колонок. Налицо несоответствие оценки результату. На *рис. 4* приведен пример такого документа. Таблица неверно разбита на несколько блоков. Смысл данных потерян.

Чтобы учитывать такие ошибки, предлагается использовать метод «Одна строка – одна ошибка». То есть, ошибкой считается каждая неправильно проанализированная строка текста. Если в одной ячейке таблицы три строки текста, то при потере ячейки такой таблицы предложенный метод оценки считает, что анализатор совершил три ошибки. Оценкой анализа будет являться сумма ошибок анализатора. Чем больше строк проанализировано правильно, тем меньше это число.


BAN – IBANESTO.COM SPA

Address:
M. Francisco Javier VALERA MARTIN, Mesena 80 – 8a Planta, E-28033 MADRID, SPAIN
Tel.: (+34) 91-338.10.57, Fax.: (+34) 91-338.31.37
E-Mail: fvaleram@eurociber.es
Internet: http://www.ibanesto.com
Organisation:
Manager: ECHAVARRI GARCIA José Miguel Team Manager: UNZUE LABIANO Eusebio. Ass. Team Manager: JAIMERENA LAURNAGARAY José Luis. Ass. Team Manager: GALILEA ZURBANO Alfonso. Public Relations: LAFARGUE Francis. Logistics Manager: VALERA MARTIN Francisco Javier

Name	Nat.	2001	2000	1999	Born
ARRIETA LUJAMBIO José Luis	SPA	BAN	BAN	BAN	710616
BARANOWSKI Dariusz	POL	BAN	BAN	BAN	720622
BARBOSA Candido	POR	BAN	BAN	BAN	741231
BENITO GUERRERO Alberto	SPA	BAN	BAN	-	750303
BLANCO GIL Santiago	SPA	BAN	VIT	VIT	740613
BROZYNA Thomas	POL	BAN	BAN	MRO	700919
BRUSEGHIN Marzio	ITA	BAN	BAN	BAN	740615
DOMINGUEZ DOMINGUEZ Juan Carlos	SPA	BAN	VIT	VIT	710413
GARCIA ACOSTA José Vicente	SPA	BAN	BAN	BAN	720804
GARCIA QUESADA Adolfo (neo)	SPA	BAN	-	-	790927
GIL PEREZ Koldo (neo)	SPA	BAN	-	-	780116
JIMENEZ SANCHEZ Eladio	SPA	BAN	BAN	BAN	760310
JIMENEZ SASTRE José Maria	SPA	BAN	BAN	BAN	710206
LASTRAS GARCIA Pablo	SPA	BAN	BAN	BAN	760120
LATASA LASA David	SPA	BAN	BAN	BAN	740214
MANCEBO PEREZ Francisco	SPA	BAN	BAN	BAN	760309
MENCHOV Denis	RUS	BAN	BAN	-	780125
MERCADO MARTIN Juan Miguel	SPA	BAN	VIT	VIT	780708
NAVAS CHICA David	SPA	BAN	BAN	BAN	740610
ODRIOZOLA MUGARZA Jon	SPA	BAN	BAN	BAN	701226
OSA EIZAGUIRRE Aitor	SPA	BAN	BAN	BAN	730909
OSA EIZAGUIRRE Unai	SPA	BAN	BAN	BAN	750612
PASCUAL RODRIGUEZ Javier	SPA	BAN	KEL	KEL	711114
PIEPOLI Leonardo	ITA	BAN	BAN	BAN	710929
PLAZA MOLINA Ruben (neo)	SPA	BAN	-	-	800229
ZANDIO ECHAIDE Xabier (neo)	SPA	BAN	-	-	770317

C.A – CREDIT AGRICOLE FRA

Address:
Vélo Club de Paris, 59 Avenue Simone, F-91800 BRUNOY, FRANCE
Tel.: (+33) 1-60.46.31.90, Fax.: (+33) 1-60.46.32.05
E-Mail: velocluhdeparis@wanadoo.fr
Organisation:
Team Manager: LEGEAY Roger. Ass. Team Manager: ROUX Denis. Ass. Team Manager: BEUCHERIE Serge. Ass. Team Manager: LAURENT Michel

Name	Nat.	2001	2000	1999	Born

Рис. 4.

Оценка основана на том, что распознавание текста проходит построчно. Таким образом, если анализ теряет строку, выделяет её не целиком или ошибочно объединяет с другой строкой, то распознавание будет проведено неверно для того же числа строк. Оценка «одна строка – одна ошибка» ориентирована на правильное распознавание максимального количества строк текста.

Однако данный метод имеет недостатки. Рассмотрим пример *(рис. 5)*. Данные в документе организованы в табличном виде, в ячейках, но таблицей этот документ не является. Нет единого признака у каждого столбца или у каждой строки таблицы. Это просто набор отдельных записей.

Анализатор OCR-системы ошибается и принимает данный документ за таблицу. Несмотря на то, что все строки текста внутри ячеек будут распознаны правильно, такой анализ согласно оценке будет иметь очень низкую оценку. Чтобы избежать это, можно усложнять оценку, учитывая отдельно строки, отдельно контекст, в котором они распознаны правильно или неправильно (таблица или просто текст), однако усложнять оценку нежелательно. Для простоты откажемся от документов такого вида и исключим их из оценочного пакета. Это - нераспространенная форма документа, а составлять оценку на нестандартных формах неразумно.

Итак, рассмотренный метод даёт разумные оценки работе анализаторов, но следует исключить из пробного пакета некоторые документы нестандартной формы.

Рис. 5

При сравнении OCR-систем FineReader 8 и FineReader 9 было подсчитано количество ошибок, согласно методу «Одна строка – одна ошибка». Данные приведены *в табл. 2.*

Таблица 2.

OCR-система	Количество ошибок, приводящих к неверному распознаванию строки
FineReader 8	7682
FineReader 9	4398

FineReader 9 сделал на 3284 ошибки меньше (т.е. правильно распознал на 3284 строк больше). От общего количества ошибок сделанных системой FineReader 8 это составляет 42,75%. На основании этих данных можно судить об улучшении качества анализа OCR-системы FineReader 9 по сравнению с предыдущей версией.

Выводы

В работе типы ошибок анализа были формализованы с помощью унифицированной системы разметки документа. Ошибки анализа делятся на 10 типов. Эти ошибки разнородны, их нельзя складывать для получения оценки вида «Количество ошибок». Поэтому предложены два типа оценки.

Первый тип оценки – сравнение количества ошибок определенного типа. Эта дифференциальная оценка, позволяет сравнивать 2 системы OCR. Анализатор системы FineReader 8 лучше справится с задачей поиска всех строк текста в документе. Но по количеству других ошибок эта система явно уступает системе FineReader 9.

Второй тип оценки – направлен на улучшение распознавания текста. Эта оценка точнее, поскольку она учитывает все неверно распознанные строки. Однако для этой оценки нужно осторожнее выбирать документы для проверочного пакета документов.

По количеству строк, неверно проанализированных двумя системами, можно сказать, что FineReader 9 лучше предыдущей версии на 43%.

Таким образом, FineReader 9 часто проводит анализ документа качественнее, чем 8-я версия системы, что согласуется с субъективной оценкой качества анализа. Предложенные оценки действительно могут использоваться для оценки блока анализа. ■

Литература

1. Арлазаров В.Л., Славин О.А. «Алгоритмы распознавания и технологии ввода текстов в ЭВМ». — Информационные технологии и вычислительные системы № 1, 1996
2. Славин О.А., Федоров Г.О. «Вопросы распознавания текста, оцифрованного с помощью видеокамер». ftp://ftp.dol.ru/pub/users/cgntv/download/sbornic/sbornic3/FEDOROV.DOC
3. Н. Е. Бузикашвили «Выделение и представление картинок на немонохромных изображениях». ftp://ftp.dol.ru/pub/users/cgntv/download/sbornic/sbornic1/buzik2.doc
4. Владимир Вежневец «Оценка качества работы классификаторов» http://cgm.graphicon.ru/content/view/106/60/
5. Kazem Taghva, Julie Borsack, Steven Lumos, Allen Condit «A comparison of automatic and manual zoning».— International Journal on Document Analysis and Recognition, Volume 6, Number 4 / April, 2003
6. Дуда Р., Харт П. Распознавание образов и анализ сцен. М., Мир, 1976
7. Коулмен Г. Б., Эндрюс Х. С. Сегментация изображений при помощи автоматической классификации. ТИИЭР, 1979, т. 67, №5, с. 39-49