Cognitive PDF/A — технология оцифровки текстовых документов для публикации в Интернете и долговременного архивного хранения *

С. А. Усилин[1], Д. П. Николаев[2], В. В. Постников[3]

[1] *ООО «Когнитивные технологии»*,
Россия, 117312 Москва, пр. 60-летия Октября, 9
[2] *Институт проблем передачи информации им. А. А. Харкевича РАН*,
Россия, 127994 Москва, Большой Каретный пер., 19, стр. 1
[3] *Институт системного анализа Российской академии наук*,
Россия, 117312 Москва, пр. 60-летия Октября, 9

В работе рассматриваются вопросы перевода бумажных документов в электронных вид. Предлагается оригинальная технология оцифровки, включающая сжатие, распознавание и упаковку текстовых документов способом, пригодным для долговременного архивного хранения. Технология использует отделение цветных элементов с помощью гистограммы насыщенности, выделение текстовых блоков с опорой на преобразование Хафа и морфологическую фильтрацию, а также методы оптического распознавания.

Введение

Увеличение объемов информации во всем мире и бурное развитие информационных технологий привело к повсеместному использованию электронных документов. Электронные версии обладают целым рядом преимуществ. Во-первых, решается проблема оперативного доступа — поиск необходимой информации, копирование, распечатка занимают секунды. Во-вторых, цифровые документы не разрушаются со временем, не ухудшают своих пользовательских качеств и могут храниться практи-

* Работа выполнена при финансовой поддержке РФФИ (грант №09–07–00473-а).

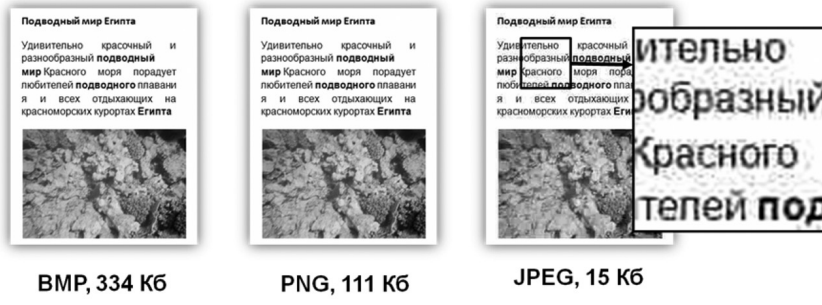


Рис. 1. Применение классических способов сжатия к изображению документа

чески вечно. В-третьих, решается проблема целостности: при правильной организации электронного архива несанкционированное удаление, фальсификация или неумышленная модификация информации невозможны. Помимо перечисленных достоинств, электронные архивы обладают рядом неявных, но удобных с практической точки зрения преимуществ: многопользовательский режим и дистанционный доступ, что абсолютно недоступно при использовании бумажного архива.

Однако большая часть знаний на сегодняшний день до сих пор хранится в бумажном виде. Наиболее популярным способом перевода бумажных документов в электронный вид является сканирование [1, 2]. В результате создаются точные электронные копии (электронные образы) документов со всеми графическими особенностями (рисунками, пометками, подписями, печатями и т. д.). При качественном сканировании получившиеся изображения-образы зачастую оказываются достаточно большого размера. Например, документ формата А4, отсканированный в цветном режиме при разрешении 300 DPI, имеет размер порядка 25 Мб. Использование файлов таких больших размеров неэффективно в электронных архивах, поэтому все больший интерес обретают технологии сжатия получившихся электронных образов. Классические технологии сжатия изображений [3, 4] не применимы, так как в общем случае документы могут содержать как монохромный текст, так и полноцветные графические области. Алгоритмы сжатия изображений без потерь, результативные для монохромных текстов, неэффективны для полноцветной графики, в то время как сжатие с потерями демонстрирует высокие показатели для цветных изображений, однако сильно искажает текстовую информацию (рис. 1). Поэтому обычно для сжатия изображений такого типа используют комбинированный подход.

1. Структурное сжатие изображений документов

Изложим идею структурного сжатия на примере изображения страницы журнала (рис. 2). Классическая страница журнала может содержать фоновый рисунок, один или несколько текстовых блоков, графических элементов (фотографии, схемы, таблицы и пр.) и каких-то пометок. Основная идея структурного сжатия изображений такого рода заключается в выделении структурных блоков, объединении данных блоков в слои (т. е. «расслоение» изображения на текстовые, графические и прочие слои) и сжатие каждого слоя наиболее подходящим образом. Так изображение страницы журнала на рис. 2 расслаивается на четыре слоя: фон, область черного текста, область синего текста и область с фотографией. Для сохранения максимального качества текстовые слои следует сжимать алгоритмами сжатия без потерь (например, CCITT Group 4), в то время как для фотографии вполне допустимо применение методов сжатия с потерями (JPEG).

Основное место в алгоритмах структурного сжатия отводится методам расслоения исходного изображения на текстовый и графический слои. На сегодняшний день опубликовано большое количество статей, предлагающих способы отделения текстовых блоков от графики различными методами [5–9]. Несмотря на это, высокую популярность данный подход получил сравнительно недавно. Одним из примеров, полностью реализующим идею структурного сжатия, по праву можно считать формат DjVu [10].

Для сжатия цветных изображений в DjVu применяется специальная технология, разделяющая исходное изображение на три слоя: передний план, задний план и черно-белую (однобитовую) маску. Маска сохраняется с разрешением исходного файла; именно она содержит изображение текста и прочие четкие детали. Разрешение заднего плана, в котором остаются иллюстрации и текстура страницы, понижается для экономии места. Передний план содержит цветовую информацию о деталях, не попавших в задний план; его разрешение понижается еще сильнее. Затем задний

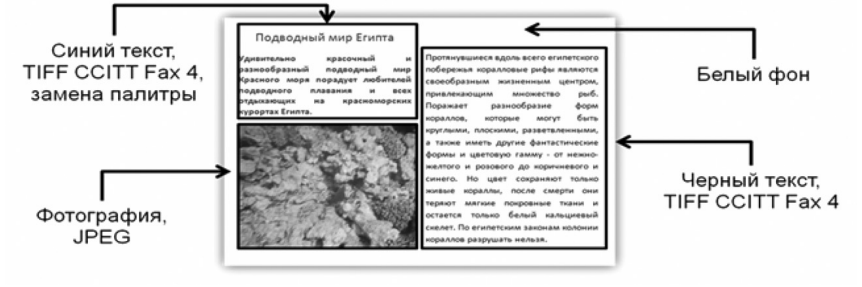


Рис. 2. Структурные блоки изображения документа

и передний планы сжимаются с помощью вейвлет-преобразования (алгоритмом IW44), а маска — алгоритмом JB2.

Несмотря на высокие коэффициенты сжатия изображений документов, DjVu обладает существенным недостатком: на сегодняшний день формат не стандартизован, что затрудняет его использование в качестве средства для создания электронных архивов. К тому же использование одинаковой схемы расслоения для всех типов документов не всегда оправдано, и даже иногда может приводить к значимому искажению документа. Дополнительно стоит отметить, что в формате полностью отсутствуют какие-либо средства обеспечения безопасности и конфиденциальности документов.

2. Технология Cognitive PDF/A

Опишем технологию Cognitive PDF/A, предназначенную для перевода бумажных документов в электронный вид, и процесс оцифровки в соответствии с предлагаемой технологией (рис. 3).

Первым этапом обработки является расслоение исходного изображения. В результате появляются два новых изображения. Первое содержит области исходного изображения, соответствующие текстовой информации (текстовый слой), а второе — графическим элементам (графический слой).

В соответствии с архитектурой алгоритма текстовый слой не должен содержать никаких лишних областей, кроме текстовых блоков. Следовательно, изображение текстового слоя может быть легко распознано без какой-либо предварительной подготовки с помощью внешних OCR-систем.

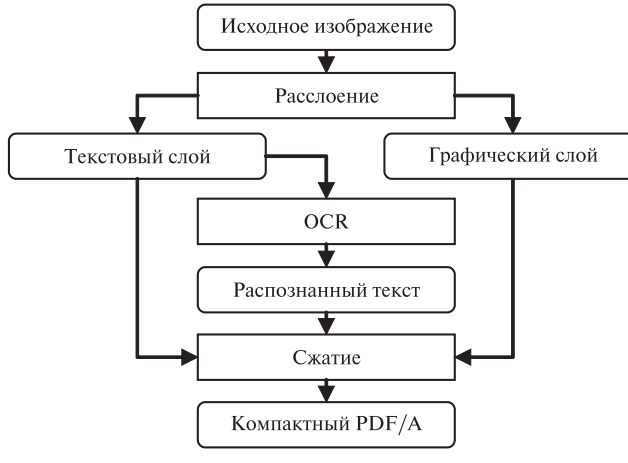


Рис. 3. Схема технологии Cognitive PDF/A

Последним действием является упаковка полученных слоев и распознанного текста в PDF/A. Графический и текстовый слои подвергаются соответствующему сжатию, а распознанный текст упаковывается таким способом, чтобы обеспечить максимальное удобство поиска и копирования информации в документе.

Таким образом, технология Cognitive PDF/A состоит из трех основных частей: расслоение исходного изображения, распознавание текстового слоя с помощью OCR-системы и компактная упаковка получившихся слоев и распознанного текста в PDF/A-файл. Рассмотрим эти части более подробно.

3. Алгоритм расслоения

Разные типы документов обладают различными особенностями. Например, для финансовых документов характерно наличие печатей, подписей и штампов, журнальные статьи могут иметь сложный многоцветный фон, в книги часто включают полноцветные графические элементы. Поэтому технологией Cognitive PDF/A предусматриваются уникальные схемы расслоения для каждого типа документа. Выбор наилучшей схемы осуществляется с помощью алгоритмов предварительной идентификации типа документа [11–13]. Далее в качестве примеров будут рассмотрены схемы расслоения для двух важных типов документов: страница книги и офисный документ.

«КВАНТ» ДЛЯ МЛАДШИХ ШКОЛЬНИКОВ

Задачи

Задачи 1 — 4 собраны из старых номеров «Кванта». Их решали младшие школьники, которые уже стали младшими и старшими научными сотрудниками.

1. Восьмиклассники построены в шеренгу. Перед каждым из них стоит семиклассник, который ниже его ростом. Докажите, что если перенги семиклассников и восьмиклассников построить по росту, то по-прежнему каждый восьмиклассник будет выше стоящего перед ним семиклассника.

(«Квант» №8 за 1974 г.)

2. Мастер спорта Седов, кандидат в мастера Чернов и перворазрядник Рыжов встретились в клубе перед началом турнира.

— Обратите внимание, — заметил черноволосый, — один из нас седой, другой рыжий, а третий черноволосый. Но ни у кого цвет волос не соответствует фамилии. Забавно, не правда ли?

— Ты прав, — подтвердил мастер.

Какого цвета волосы у кандидата в мастера?

(«Квант» №10 за 1976 г.)

3. Десять человек пришли в гости в галошах. Уходили они по одному, и каждый надевал произвольную пару галош, в которую он мог влезть (т.е. не меньшего размера, чем его собственная). Какое наибольшее число людей не смогло надеть галоши?

(«Квант» №5 за 1978 г.)

4. Из спичек сложено слово «ТОЛЯ»:

ТОЛЯ

Переложите ровно одну спичку так, чтобы получилось женское имя.

(«Квант» №2 за 1976 г.)

5. На каждой горизонтали и каждой вертикали шахматной доски стоит нечетное число фигур. Докажите, что на черных клетках шахматной доски стоит четное число фигур.

В.Произволов

Иллюстрации Д.Гришанова

Рис. 4. Изображение страницы книги

а)

б)

Рис. 5. Изображение страницы книги после бинаризации (а); после применения морфологической операции «размыкание» (б)

Обычно страница книги содержит черный текст на белом фоне и, возможно, графические элементы: рисунки, схемы, графики и пр. (рис. 4). Обычно в книгах области текста и графики не пересекаются. Еще одной ключевой особенностью верстки книг является использование шрифтов близких линейных размеров. Опираясь на эти характерные черты, построим схему расслоения изображения страницы книги.

Шаг 1. Бинаризуем исходное изображение [14, 15], тем самым преобразуем его в монохромный вид (рис. 5 *а*). Так как изображение в основном содержало черный текст на белом фоне, то процесс бинаризации не должен сильно сказаться на областях, содержащих текстовую информацию.

Шаг 2. С помощью морфологической фильтрации [4, 16] «сольем» слова в единые компоненты связности. Обозначим w и h характерные ширину и высоту символов соответственно. Заметим также, что расстояние между буквами в слове сравнимо с толщиной штриха символа, а расстояние между словами близко ширине символа. Поэтому «склеим» каждое слово

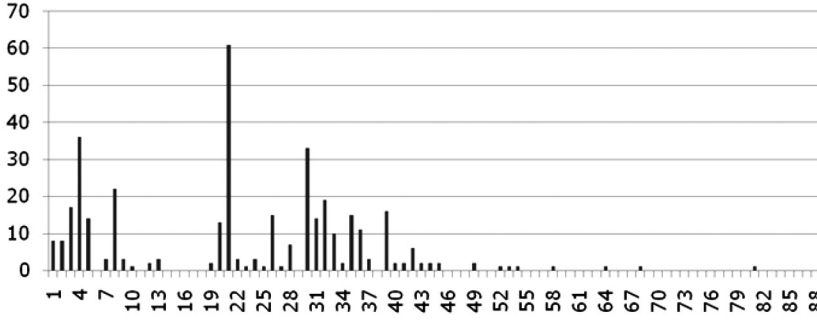
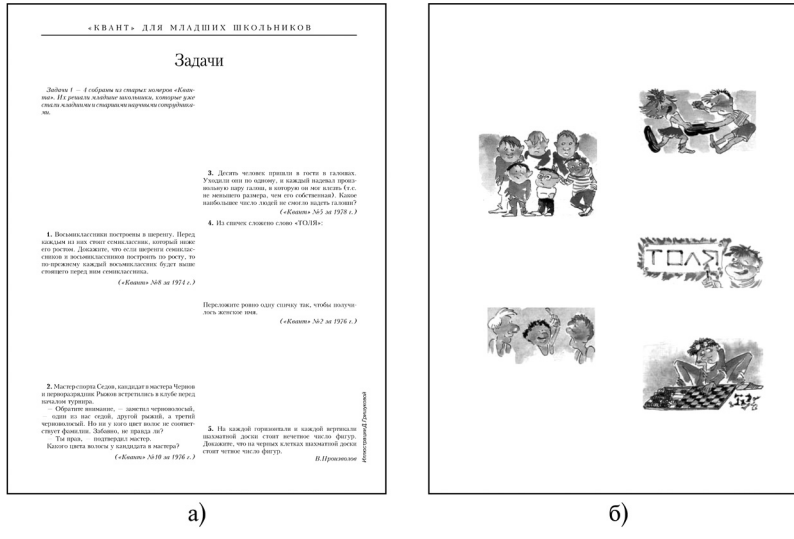


Рис. 6. Гистограмма высот компонент связности



а) б)

Рис. 7. Текстовый (а) и графический (б) слои изображения страницы книги

в отдельную компоненту связности, выполнив размыкание с окном $w \times 1$ (рис. 5 *б*). Прямое вычисление морфологического фильтра имеет линейную сложность от размера окна. Поэтому для ускорения используется алгоритм Ван Херка [17], вычисляющий морфологические операции с прямоугольным примитивом за время, не зависящее от размеров окна.

Шаг 3. Построим гистограмму высот полученных компонент связности (рис. 6). Так как весь текст на странице напечатан примерно одинаковым по размеру шрифтом, то компоненты связности, соответствующие

S-SHINA.RU
ваш магазин дисков и шин
тел. (495) 730-57-34, 223-22-57

Расходная накладная № 1875 от 28 августа 2008 09:37:00
Действительна только в случае оплаты в кассе продавца в день выписки до конца рабочего дня.

2 100001 643911

Поставщик: Общество с ограниченной ответственностью "Вилс Авто"
Покупатель: Частное лицо (Клиент-"Варшавка")
№ машины:
Ответственный менеджер: Салахетдинов Руслан Гафаевич
Подразделение: Варшавское шоссе, д. 46

№	Артикул	Товар	Количество		Цена	Сумма без скидки	Скидка	Сумма	Номер ГТД / Страна происхождения
1	A3337k	А/Диск 5x114.3 6.5x15 ET38 d67,1 КиК Алатау	4	шт	2 332,00	9 328,00	652,96	8 675,04	
2	ру708113	А/Шина 195/65 R15 91T Goodyear UltraGrip Extreme	4	шт	3 317,00	13 268,00	928,76	12 339,24	
		общее количество:	8						
					Итого:	22 596,00	1 581,72	21 014,28	
					В том числе НДС:			3 205,57	

Всего наименований 2, на сумму 21 014,28 руб.
Двадцать одна тысяча четырнадцать рублей 28 копеек

Заказ оформил
Отпустил
(Мартынов Михаил)
Получил

Уважаемые клиенты! При покупке в нашем магазине комплекта шин или дисков Вы получаете скидку размером 10% на услуги Шиномонтажа и Автосервиса!
Не забывайте спрашивать купоны на скидку у менеджеров.

ОПЛАЧЕНО

Рис. 8. Пример изображения финансового документа

словам, образуют на гистограмме один или несколько четко выраженных максимумов. Поэтому, анализируя гистограмму, можно вычислить характерный размер шрифта h_{font}, которым набран текст на странице, и, соответственно, выделить область на изображении, соответствующую текстовой информации (области, соответствующие компонентам связности с высотой порядка h_{font}).

Зная область расположения текста на исходном документе, построим маску расслоения, после чего применим ее для получения графического и текстовых слоев (рис. 7).

Поскольку для выделения текстовых блоков используются быстрые алгоритмы морфологической фильтрации с прямоугольным окном, то очень важно, чтобы текстовые блоки были выровнены относительно осей изображения. Поэтому перед морфологией выполняется «выравнивание» изображения с опорой на быстрое преобразование Хафа [18].

Для цветных изображений финансовых документов (счетов-фактур, квитанций, договоров и пр.) не характерны перечисленные выше особенности изображения страницы книги, так как графические элементы (печати, подписи, рукописные пометки) часто накладываются на текстовые блоки (рис. 8). Следовательно, использовать для расслоения вышеописанный алгоритм неразумно. Построим схему расслоения, опираясь на цветовые характеристики изображения. Цветовая насыщенность [4] черного текста и белого фона близка к нулю, в то время как для синих печатей и подпи-

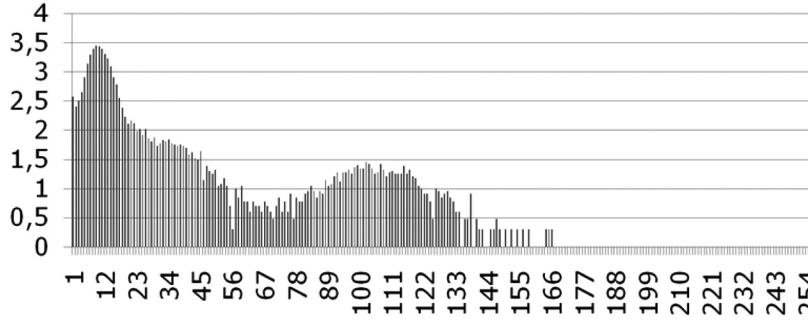


Рис. 9. Гистограмма цветовой насыщенности

сей это значение велико. Принимая во внимание это свойство, построим следующую схему расслоения.

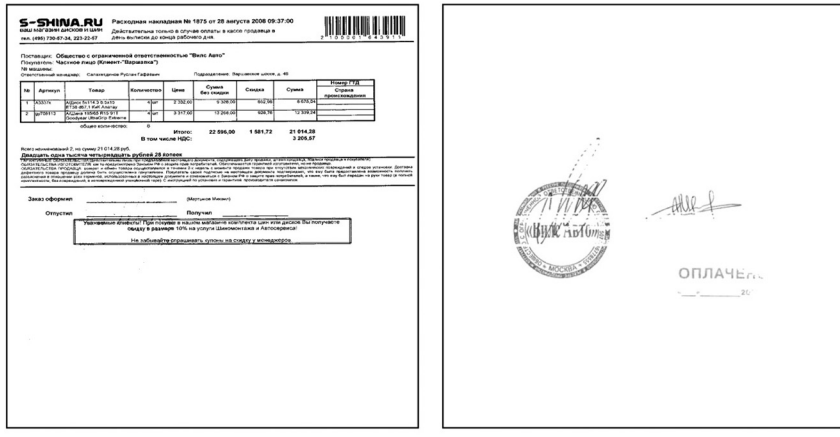
Шаг 1. Построим гистограмму цветовой насыщенности (рис. 9), т. е. зависимость $y = \log N_x$, где N_x — количество пикселей изображения, насыщенность которых равна x.

Шаг 2. Заметим, что на гистограмме четко выделяются два класса: первый сформирован пикселями с малыми значениями цветовой насыщенности, второй — с большими значениями. Пиксели из первого класса составляют области изображения, соответствующие фону и черному тексту, из второго — графическую часть изображения. Найдем порог разделения двух классов t^* методом Отсу [15].

Шаг 3. Если величина порога t^* слишком мала (меньше, чем некоторое минимальное значение $t_{\min}$), то исходное изображение содержит только черный текст и не содержит каких-либо цветных графических элементов. В противном случае расслоим исходное изображение следующим образом: пиксель исходного изображения (x, y) принадлежит текстовому слою (рис. 10 *а*), если значение его цветовой насыщенности меньше порогового $s(x, y) < t^*$; иначе — пиксель (x, y) принадлежит графическому слою (рис. 10 *б*).

4. Распознавание текстового слоя

В результате расслоения исходное изображение разделяется на два новых: изображение, содержащее только текстовую информацию, и изображение, содержащее только графические элементы. Изображение, соответст-



а) б)

Рис. 10. Текстовый (а) и графический (б) слои изображения финансового документа

вующее текстовому слою, может быть легко распознано без какой-либо предварительной подготовки с помощью внешних OCR-систем.

В программной реализации технологии Cognitive PDF/A в качестве OCR-модуля выбрана система оптического распознавания текстов OCR CuneiForm [19], предназначенная для распознавания любых полиграфических шрифтов, получаемых с принтеров, за исключением декоративных и рукописных.

Система снабжена многими современными алгоритмами и технологиями, повышающими качество распознавания, например адаптивное распознавание и алгоритмы распознавания, использующие нейронные сети. Адаптивное распознавание [20] — метод, основанный на комбинации двух видов алгоритмов распознавания печатных символов: шрифтового (multifont) и шрифтонезависимого (omnifont). Система генерирует внутренний шрифт для каждого вводимого документа, основываясь на хорошо пропечатанных символах, т. е. используется динамическая настройка (адаптация) на конкретные входные символы. Таким образом, метод совмещает универсальность и технологичность безшрифтового подхода и высокую точность распознавания шрифтового, что позволяет кардинальным образом повысить качество распознавания. Алгоритмы, использующие нейронные сети для распознавания символов, строятся следующим образом [21]. Поступающее на распознавание изображение символа (растр) приводится к некоторому стандартному размеру (нормализуется). Значения яркости в узлах нормализованного растра используются в качестве входных параметров нейронной сети. Число выходных параметров

нейронной сети равняется числу распознаваемых символов. Результатом распознавания является символ, которому соответствует наибольшее из значений выходного вектора нейронной сети. В систему встроены специальные алгоритмы для распознавания текста с матричного принтера, плохих ксерокопий факсов.

Первоначально система CuneiForm была разработана компанией Cognitive Technologies как коммерческий продукт и поставлялась с некоторыми моделями сканеров. Однако в 2008 году было анонсировано открытие исходных текстов программы в соответствии с лицензией BSD, что позволяет использовать данное программное обеспечение в сторонних продуктах.

5. Сжатие и упаковка в формат PDF/A

Полученные в результате расслоения текстовый и графический слои, а также распознанный текст сохраняются в формате PDF/A. Данный формат является стандартом ISO 19005–1:2005, базируется на описании стандарта PDF версии 1.4 от Adobe Systems Inc. и предназначен специально для долгосрочного архивного хранения электронных документов [22]. Несмотря на то что PDF/A является подмножеством формата PDF, существует ряд различий, обусловленных требованиями, предъявляемыми к PDF/A как к формату долгосрочного хранения электронных документов. Так, например, обязательным для PDF/A является:

- внедрение всех используемых шрифтов, в том числе шрифтов из списка «стандартных для PDF»;
- внедрение цветового профиля (если PDF/A-файл содержит изображения) — файла, в котором содержится информация о том, как выводное устройство (монитор, принтер и пр.) должно передать цвет. Важным является тот факт, что включаемый цветовой профиль должен быть аппаратно-независимым;
- обязательное наличие метаданных с указанием версии используемого формата, заголовка документа, списка авторов, краткого описания, даты создания и последней модификации файла документа, а также ключевых слова для осуществления поиска. Спецификацией PDF/A также оговорен формат представления метаданных — Adobe Extensible Metadata Platform (XMP) [23].

Для увеличения коэффициента сжатия графический и текстовый слои сжимаются разными способами. В силу специфики содержимого графический слой приводится к разрешению 100 DPI и кодируется алгоритмом JPEG. Текстовый слой заключает в себе основную информацию документа, следовательно, текстовый слой сохраняется в исходном разрешении, а для кодирования используются алгоритм сжатия без потерь CCITT Group 4.

Заключение

Для определения эффективности технологии было выбрано несколько изображений различного типа, которые были сжаты алгоритмами JPEG (уровень компрессии, при котором сохраняется читабельность), DjVu и Cognitive PDF/A.

По результатам сравнения видно (см. табл. 1), что описанная технология по степени сжатия на порядок обходит JPEG, однако существенно проигрывает DjVu. Такую значительную разницу в размере можно объяснить тем, что файл в формате PDF/A помимо полезной информации (собственно изображений и распознанного текста) содержит также вспомогательные данные, необходимые для долгосрочного хранения. К тому же формат DjVu использует более совершенный алгоритм сжатия с потерями (IW44), основанный на вейвлет-преобразованиях, который не поддерживается настоящей версией формата PDF/A. Однако несмотря на меньший размер файлов DjVu качество сжатия офисных документов довольно низкое, что особенно проявляется в районе печатей и подписей.

Изложенный в работе алгоритм оцифровки текстовых документов реализован программно и внедрен в систему Cognitive Forms [2] в качестве средства архивирования документов. Данная программная подсистема была промышленно внедрена и используется для массового ввода документов. В проекте, реализованном для ОАО «Магнитогорский металлургический комбинат», разработанный алгоритм сжатия применяется для обработки бухгалтерских документов (счета-фактуры, накладные, акты, договора). В проекте, реализованном для ООО «РусФинанс», алгоритмы сжатия используются для обработки изображений документов, необходимых для выдачи кредитов (заявления на выдачу кредита, свидетельства ИНН, ксерокопии паспорта и водительского удостоверения и пр.).

Таблица 1

Сравнение различных технологий сжатия

Изображение	JPEG	DjVu	Cognitive PDF/A
Рис. 11 *а*	718 KB	50 KB	94 KB
Рис. 11 *б*	373 KB	29 KB	51 KB
Рис. 11 *в*	642 KB	23 KB	52 KB
Рис. 11 *г*	120 KB	20 KB	40 KB
Рис. 11 *д*	292 KB	17 KB	29 KB

«КВАНТ» ДЛЯ МЛАДШИХ ШКОЛЬНИКОВ

Задачи

Задачи 1 — 4 собраны из старых номеров «Кванта». Их решали младшие школьники, которые уже стали младшими и старшими научными сотрудниками.

1. Восьмиклассники построены в шеренгу. Перед каждым из них стоит семиклассник, который ниже его ростом. Докажите, что если шеренги семиклассников и восьмиклассников построить по росту, то по-прежнему каждый восьмиклассник будет выше стоящего перед ним семиклассника.

(«Квант» №8 за 1974 г.)

2. Мастер спорта Седов, кандидат в мастера Чернов и перворазрядник Рыжов встретились в клубе перед началом турнира.

— Обратите внимание, — заметил черноволосый, — один из нас седой, другой рыжий, а третий черноволосый. Но ни у кого цвет волос не соответствует фамилии. Забавно, не правда ли?

— Ты прав, — подтвердил мастер.

Какого цвета волосы у кандидата в мастера?

(«Квант» №10 за 1976 г.)

3. Десять человек пришли в гости в галошах. Уходили они по одному, и каждый надевал произвольную пару галош, в которую он мог влезть (т.е. не меньшего размера, чем его собственная). Какое наибольшее число людей не смогло надеть галоши?

(«Квант» №5 за 1978 г.)

4. Из спичек сложено слово «ТОЛЯ»:

ТОЛЯ

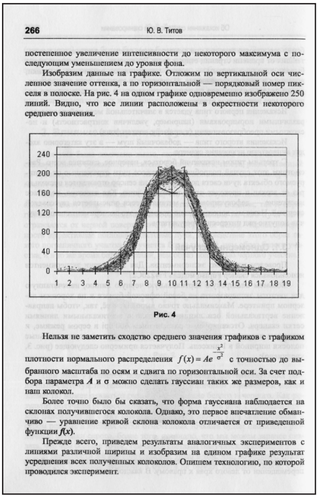
Переложите ровно одну спичку так, чтобы получилось женское имя.

(«Квант» №2 за 1976 г.)

5. На каждой горизонтали и каждой вертикали шахматной доски стоит нечетное число фигур. Докажите, что на черных клетках шахматной доски стоит четное число фигур.

В.Произволов

а)



266 Ю. В. Титов

постепенное увеличение интенсивности до некоторого максимума с последующим уменьшением до уровня фона.

Изобразим данные на графике. Отложим по вертикальной оси численное значение оттенка, а по горизонтальной — порядковый номер пикселя в полоске. На рис. 4 на одном графике одновременно изображено 250 линий. Видно, что все линии расположены в окрестности некоторого среднего значения.

Рис. 4

Нельзя не заметить сходство среднего значения графиков с графиком плотности нормального распределения $f(x) = Ae^{-\frac{x^2}{\sigma^2}}$ с точностью до выбранного масштаба по осям и сдвига по горизонтальной оси. За счет подбора параметра A и σ можно сделать гауссиан таких же размеров, как и наш колокол.

Более точно было бы сказать, что форма гауссиана наблюдается на склонах получившегося колокола. Однако, это первое впечатление обманчиво — уравнение кривой склона колокола отличается от приведенной функции $f(x)$.

Прежде всего, приведем результаты аналогичных экспериментов с линиями различной ширины и изобразим на едином графике результат усреднения всех полученных колоколов. Опишем технологию, по которой проводился эксперимент.

б)

СБ7970/1279
25.08.2008 10:58 122
1 0191
044583001
40503810600001009079
7707020131
06319356940
31930201010010000130
45286585000
11000.00 0.00
КОМИССИЯ 330.00
ИТОГО: 11330.00

Квитанция

Кассир

УФК по г. Москве (ИППИ РАН л/с 06319356940) КПП 7 7 0 7 0 1 0 0 1
(наименование получателя платежа)
7 7 0 7 0 2 0 1 3 1 — 45286585000
(ИНН налогового органа) и его сокращенное наименование Код ОКАТО
40503810600001009079 в Отделении 1 Московского ГТУ Банка России г.Москва 705
(номер счета получателя платежа) (наименование банка)
БИК: 044583001 Кор./сч.:
Оргвзнос за участие в конференции ИТиС-2008 31930201010010000130
(наименование платежа) (код бюджетной классификации КБК)
Плательщик (Ф.И.О.) Усилин Сергей Александрович
Адрес плательщика г. Сергиев Посад, ул. Октябрьская, 1-6
ИНН плательщика № л/с плательщика
Сумма 11000 руб. 00 коп.
Плательщик (подпись) Дата: 25 августа 2008 г.
* или иной государственный орган исполнительной власти

в)

СМЕТА № [illegible] от [illegible] 2007г.

на производство ремонтных работ

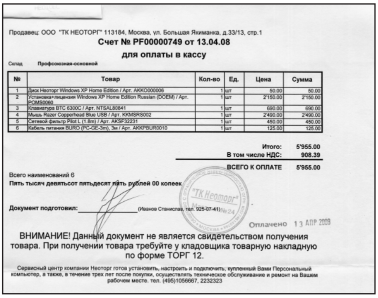
Настоящая смета составлена [illegible] в лице Ген. [illegible] Борисовича, в дальнейшем именуемого «Подрядчик» и ______________________

В дальнейшем именуемого «Заказчик», в том, что «Подрядчик» обязуется выполнить ремонтные работы на объекту по адресу______________________ В полном объеме и по срокам указанном в договоре №_____от___________2007г. По ценам указанным в настоящей смете. Смета составлена в 2-х экземплярах с приложениями на__8__листах. Всякие изменения в смете после ее подписания недопустимы. Изменения происходящие в ходе работ вносятся в приложение к смете (лист «Приложения») м подписываются сторонами. Данная смета согласованна сторонами_______________2007г. Общая сметная стоимость проводимых работ Составляет [illegible]

«ЗАКАЗЧИК» **«ПОДРЯДЧИК»**

Ген.директор [illegible]

г)



Продавец: ООО "ТК НЕОТОРГ" 113184, Москва, ул. Большая Якиманка, д.33/13, стр.1

Счет № PF00000749 от 13.04.08

для оплаты в кассу

Склад **Профсоюзная-основной**

№	Товар	Кол-во	Ед.	Цена	Сумма
1	Диск Неоторг Windows XP Home Edition / Арт. AKKO000006	1	шт	50.00	50.00
2	Установка+лицензия Windows XP Home Edition Russian (DOEM) / Арт. POMS0060	1	шт	2'150.00	2'150.00
3	Клавиатура BTC 6300C / Арт. NTSAL80841	1	шт	690.00	690.00
4	Мышь Razer Copperhead Blue USB / Арт. KKMSRS002	1	шт	2'490.00	2'490.00
5	Сетевой фильтр Pilot L (1.8m) / Арт. AKSF32231	1	шт	450.00	450.00
6	Кабель питания BURO (PC-GE-3m), 3м / Арт. AKKPBUR0010	1	шт	125.00	125.00

Итого: 5'955.00

В том числе НДС: 908.39

ВСЕГО К ОПЛАТЕ 5'955.00

Всего наименований 6

Пять тысяч девятьсот пятьдесят пять рублей 00 копеек

Документ подготовил:________________(Иванов Станислав, тел. 925-07-41)

"ТК Неоторг" №24

Оплачено 13 АПР 2008

ВНИМАНИЕ! Данный документ не является свидетельством получения товара. При получении товара требуйте у кладовщика товарную накладную по форме ТОРГ 12.

Сервисный центр компании Неоторг готов установить, настроить и подключить, купленный Вами Персональный компьютер, а также, в течение трех лет после покупки, осуществлять техническое обслуживание и ремонт на Вашем рабочем месте. тел. (495)1056667, 2232323

д)

Рис. 11. Примеры изображений различного типа: страница книги с цветными иллюстрациями (а); страница книги с графикой (б); квитанция (в); финансовый документ с насыщенной печатью (г); финансовый документ с тусклой печатью и штампом (д)

Литература

1. *Арлазаров В. Л., Славин О. А.* Алгоритмы распознавания и технологии ввода текстов в ЭВМ // Информационные технологии и вычислительные системы. 1996. Т. 6. № 1. С. 48–54.
2. *Арлазаров В. В., Постников В. В., Шоломов Д. Л.* Cognitive Forms — система массового ввода структурированных документов // Управление информационными потоками. М: URSS, 2002. С. 37–49.
3. *Ватолин Д., Ратушняк А., Смирнов М., Юкин В.* Методы сжатия данных. Алгоритмы сжатия изображений. М.: Диалог-МИФИ, 2002. 99 с.
4. *Гонсалес Р., Вудс Р.* Цифровая обработка изображений. М.: Техносфера, 2005. 1072 с.
5. *Fletcher L. A., Kasturi R.* A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1988. Vol. 10. № 6. С. 910–918.
6. *Hasan Y., Karam L.* Morphological text extraction from images // IEEE Transactions on In Image Processing. 2000. Vol. 9. № 11. С. 1978–1983.
7. *Hirata N. S. T., Barbera J., Terada R.* Text segmentation by automatically designed morphological operators // Computer Graphics and Image Processing. 2000. С. 284–291.
8. *Yuan Q., Tan C. L.* Text extraction from gray scale document images using edgeinformation // IEEE Document Analysis and Recognition Processing. 2001. С. 302–306.
9. *Cao R.* Separation of overlapping text from graphics // IEEE Document Analysis and Recognition Processing. 2001. С. 44–48.
10. Lizardtech DjVu Reference. http://www.lizardtech.com
11. *Postnikov V. V.* Flexible Forms Identification // Proceedings of the 5th German-Russian Workshop on Pattern Recognition and Image Understanding (GRWS98). Hamburg, 1999.
12. *Усилин С. А., Николаев Д. П., Постников В. В.* Быстрый алгоритм совмещения изображений документов в произвольной геометрической модели // Труды конференции «Информационные технологии и системы». 2008. С. 471–477.
13. *Безматерных П. В., Николаев Д. П., Постников В. В.* Метод идентификации типа документа по структуре проекций его изображения на координатные оси // Труды конференции «Информационные технологии и системы». 2008. С. 498–501.
14. *Nikolaev D. P.* Segmentation-based binarization method for color document images // Proceedings of 6th Open Russian-German Workshop on Pattern Recognition and Image Understanding. 2003. С. 190–193.
15. *Trier O. D., Taxt T.* Evaluation of binarization methods for document images // IEEE Transactions on Pattern Analysis and Machine Intelligence. 1995. Vol. 17. № 3. С. 312–315.
16. *Куроптев А. В., Николаев Д. П., Постников В. В., Усилин С. А.* Выделение графических примитивов и текстовых блоков на изображениях документов с помощью морфологических операций // Труды 51-й научной конференции МФТИ. Современные проблемы фундаментальных и прикладных наук. Часть 9. Инновации и высокие технологии. М.: МФТИ, 2008. С. 29–31.

17. *Herk M. van.* A Fast Algorithm for Local Minimum and Maximum Filters on Rectangular and Octagonal Kernels // Pattern Recognition Letters. 1992. C. 517–521.
18. *Nikolaev D. P., Karpenko S. M., Nikolaev I. P., Nikolayev P. P.* Hough Transform: Underestimated Tool in the Computer Vision Field // 22st Europ. Conf. Model. Simulat. ECMS 2008. Nicosia, Cyprus, 2008. P. 238–243.
19. OCR CuneiForm. http://www.cuneiform.ru
20. *Арлазаров В. Л., Астахов А. Д., Троянкер В. В., Котович Н. В.* Адаптивное распознавание символов // Интеллектуальные технологии ввода и обработки информации. М.: URSS, 1998. С. 39–56.
21. *Мисюрёв А. В.* Использование искусственных нейронных сетей для распознавания рукопечатных символов // Интеллектуальные технологии ввода и обработки информации. 1998. С. 122–127.
22. ISO 19005–1:2005. Document management — Electronic document file format for long-term preservation. Part 1: Use of PDF 1.4 (PDF/A-1).
23. Adobe Systems Incorporated. Extensible Metadata Platform (XMP) Specification. http://www.adobe.com