Любарский М.Г.
д.ф.-м.н.,
проф.,
Национальный юридический университет Украины
им. Я. Мудрого
Кафедра информатики и вычислительной техники
г. Харьков, Украина

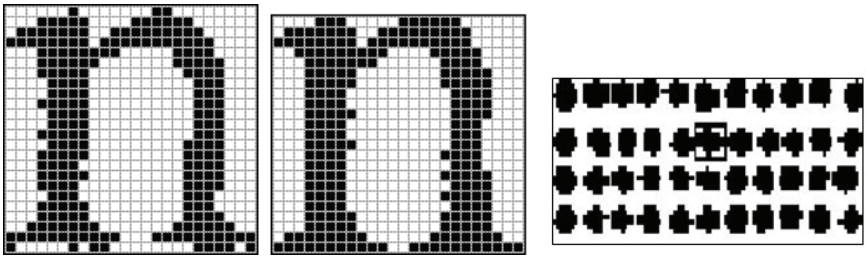
КОМПРЕССИЯ ИЗОБРАЖЕНИЯ ТЕКСТОВЫХ ДАННЫХ НА ОСНОВЕ ДЕТАЛИЗАЦИИ СИМВОЛОВ, СТАТИСТИЧЕСКОГО АНАЛИЗА И КЛАССИФИКАЦИИ

В данной работе предложен новый метод сжатия битонального изображения текста, где в качестве основных элементов обработки рассматриваются не связные символы изображения текста, а результат их детализации - вертикальные элементы строки. Разработана вероятностная модель, алгоритм статистического анализа и классификации вертикальных элементов строки. Применение представленного метода позволяет получить преимущество в степени сжатия в сравнении с алгоритмом JB2 формата DjVu около 37 % при наиболее используемом разрешении изображения текста в 300 dpi.

Современные методы сжатия, основанные на различных ортогональных преобразованиях, дают хороший результат при сжатии размытых изображений, но не эффективны для битональных изображений, тем более изображений текста, изобилующего множеством мелкими деталей – буквами, цифрами, знаками препинания. В настоящее время лучшие алгоритмы для сжатия битональных изображений текста основаны на выделении изображений символов и их классификации. Это – алгоритмы JB2 и JBIG2, используемые соответственно в широко распространённых форматах DjVu и PDF [1–4]. Степень сжатия информации с помощью методов классификации тем выше, чем меньше классов образуется при классификации и чем больше элементов в каждом классе [5-7]. В идеале при сжатии изображения страницы текста изображения каждого символа должны находиться в одном и только одном классе. Однако ни один из известных алгоритмов этому условию не удовлетворяет. Дело в шумах (случайных искажениях), возникающих при печати страницы и ее последующем сканировании. На рис. 1, а, представлены два случайно выбранные изображения буквы «n» из различных 257, входящих в изображение страницы текста формата А4, при разрешении сканирования 300 dpi.

Результатом влияния контурных шумов на символы изображения является то, что на странице не найдется ни одной пары символов «n», полностью совпадающих друг с другом. То же относится и к другим символам, даже точкам, рис. 1, *б*.

Рис.: 1. Влияние шумов на изображения символов: а – искажения символа «n»; б – искажения символа «точка»

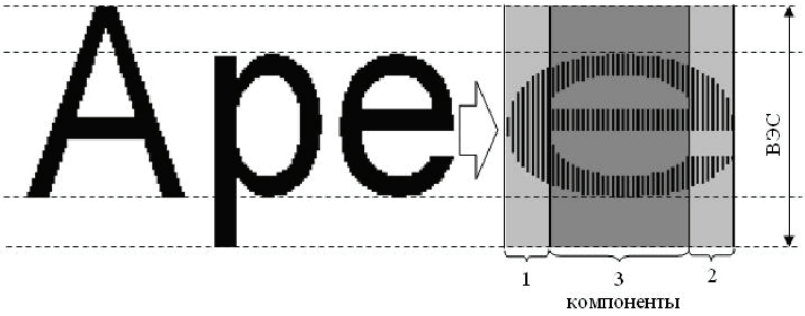


а б

Указанные недостатки алгоритмов, классифицирующих изображения символов, наводят на мысль о том, что хотя выбор изображений символов в качестве элементов изображения страницы является естественным, этот выбор не является оптимальным.

Новый подход к сжатию графических текстовых данных заключается в следующем. Если представить себе прямоугольник, охватывающий какую-либо строку, то *вертикальным элементом* этой строки будем называть пересечение прямоугольника с любой вертикальной линией шириной в один пиксель. На рис. 2 показано разбиение изображения буквы «е» на вертикальные элементы строки.

Рис.: 2. Изображение буквы «е» и составляющие его вертикальные элементы строки с различным числом компонент



Таким образом, страницу текста можно рассматривать как упорядоченную совокупность вертикальных элементов. Такое разбиение удобно тем, что все вертикальные элементы имеют один и тот же размер и их можно представлять и как двоичные числа, и как векторы с координатами 0 (черный пиксель) и 1 (белый пиксель).

Шумы печати и сканирования случайным образом искажают вертикальные элементы. Так что среди них могут быть искаженные и неискаженные элементы. Однако бессмысленно разбивать совокупность вертикальных элементов, составляющих изображение страницы, на классы тождественных или почти тождественных элементов, поскольку многие из них могут быть искажениями сразу нескольких неискаженных элементов. Более того, встречаются пары неискаженных элементов, которые совпадают с искажениями друг друга.

Имеет смысл говорить только о нечеткой классификации вертикальных элементов, то есть о вероятности того, что данный элемент есть искажение того или иного неискаженного элемента. При этом вопрос о том, является ли какой-то элемент неискаженным, тоже имеет лишь вероятностный ответ.

Таким образом, основная задача статистического анализа совокупности вертикальных элементов, представляющих текстовую страницу, ставится так: *по имеющейся на странице совокупности* X *вертикальных элементов указать минимальную наиболее правдоподобную совокупность* $C \subset X$ *неискаженных элементов, а также для каждой пары* $x \in X$ *и* $c \in C$ *найти вероятность того, что данный элемент* x *является искажением элемента* c.

После нахождения этих вероятностей легко получить правильную классификацию изображений символов, представив последние как упорядоченный набор вертикальных элементов. Грубо говоря, изображения двух символов можно отнести к одному классу, если у каждой пары вертикальных элементов, составляющих эти изображения и имеющих один и тот же порядковый номер, достаточно велика вероятность того, что они являются искажениями одного и того же вертикального элемента.

Используя отдельный этап детализации связных символов изображения текста, в виде вертикальных элементов строки и применив их нечеткую классификацию, была получена минимальная наиболее правдоподобная совокупность неискаженных элементов строки. Учитывая ограничения вероятностной модели, для каждого вертикального элемента строки была получена вероятность того, что он является искажением найденного неискаженного элемента строки.

Исследовав возможности компрессии данных представленных в виде словаря неискаженных вертикальных элементов строки и их карты размещения, была проведена классификация связных символов на основе нечеткой классификации вертикальных элементов. Формирование словаря связных символов основывалось не на сравнительном анализе геометрических форм сравниваемых символов [8], а на вероятностной оценке соответствующих вертикальных элементов строки, которые представляют состав классифицируемых связных символов.

Предложенный алгоритм представления и обработки изображения текста позволил получить достаточно высокую степень сжатия при хорошем качестве восстановленного изображения.

Сравнение с лучшим в настоящее время специальным алгоритмом сжатия изображений текста – JB2, входящим в формат DjVu, показало, что предлагаемый алгоритм сжатия изображения текста имеет преимущество в степени компрессии данных порядка 37% при обработке страницы текста изображения. Для наиболее часто используемого на практике разрешения изображения текста 300 dpi авторами были получены следующие сравнительные количественные показатели сжатия:

- в работе [8] преимущество над JB2 – 8 %;
- в работе [9] преимущество над JB2 – 25 %;
- в работе [10] преимущество над JB2 – 37 %.

Это является основной характеристикой представленного метода и раскрывает новые возможности повышения информативности представления текстовых графических данных в инженерных реализациях.

Литература:

1. Technical Papers from AT&T Labs [**Electronic Resource**] / Available at**:** http://djvuzone.org/techpapers/index.html
2. DjVu.org [**Electronic Resource**] / Available at**:** http://www.djvu.org/
3. Haffner P. DjVu: Analyzing and Compressing Scanned Documents for Internet Distribution [Text] / P. Haffner, L. Bottou, P. G. Howard, Y. LeCun // Fifth International Conference on Document Analysis and Recognition (ICDAR'99), 1999. – P. 625
4. JBIG2.com : An Introduction to JBIG2 [**Electronic Resource**] / **available at : URL :** http://jbig2.com/index.html
5. Айвазян С. А. Прикладная статистика: Классификация и снижение размерности [Текст] / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков и др. – М.: Финансы и статистика, 1989. – 607 с.

6. Иванов В. Г. Сжатие изображений на основе автоматической и нечеткой классификации фрагментов [Текст] / В. Г. Иванов, Ю. В. Ломоносов, М. Г. Любарский // Проблемы управления и информатики. – 2009. – № 1 – с. 52–63.
7. Шлезингер М. И. Математические средства обработки изображений [Текст] / М. И. Шлезингер. – Киев: Наукова думка, 1983. – 200 с.
8. Иванов В. Г. Сжатие изображения текста на основе выделения символов и их классификации [Текст] / В. Г. Иванов, М. Г. Любарский, Ю. В. Ломоносов // Проблемы управления и информатики. – 2010. – № 6. – с. 111–122.
9. Иванов В. Г. Сжатие изображения текста на основе формирования и классификации вертикальных элементов строки в графическом словаре символьных данных [Текст] / В. Г. Иванов, М. Г. Любарский, Ю. В. Ломоносов // Проблемы управления и информатики. – 2011. – № 5. – с. 98–109.
10. Иванов В. Г. Сжатие изображения текста на основе статистического анализа и классификации вертикальных элементов строки [Текст] / В. Г. Иванов, Ю. В. Ломоносов, М. Г. Любарский // Восточно-Европейский журнал передовых технологий. – 2014. - № 4/2 (70). – с. 4-15.

Луценко О.П.
Дніпропетровський національний університет імені О. Гончара
Кафедра математичного забезпечення ЕОМ
м. Дніпропетровськ, Україна
Байбуз О.Г.
д.т.н., проф.,
Дніпропетровський національний університет імені О. Гончара
Кафедра математичного забезпечення ЕОМ
м. Дніпропетровськ, Україна

ОЦІНКА ФУНКЦІЇ РИЗИКУ РОЗЛАДНАННЯ ПРОЦЕСУ КОЛИВАНЬ ВАЛЮТНИХ КУРСІВ З ЗАСТОСУВАННЯМ БАЙЕСІВСЬКОЇ ОЦІНКИ ПАРАМЕТРУ ФУНКЦІЇ УМОВНОГО РОЗПОДІЛУ

В попередніх публікаціях [1] авторами була запропонована модель стохастичної оцінки стану фінансового ринку, заснована на методах виявлення розладнання і відтворення щільності розподілу.