УДК 621.391:519.728

ИВАНОВ В.Г., ЛОМОНОСОВ Ю.В., ЛЮБАРСКИЙ М.Г.
НЮУ имени Ярослава Мудрого (Украина)

НЕЧЕТКАЯ КЛАССИФИКАЦИЯ И СЖАТИЕ ИЗОБРАЖЕНИЯ ТЕКСТА

Современные методы сжатия, основанные на различных ортогональных преобразованиях, дают хороший результат при сжатии размытых зображений. Однако, такие методы не эффективны для битональных изображений, тем более изображений текста, которые состоят из множества мелких деталей – букв, цифр, знаков препинания. В данной работе представлены последние результаты исследований по обработке изображений текста.

Введение. 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 печати и сканировании страницы, рис. 1.

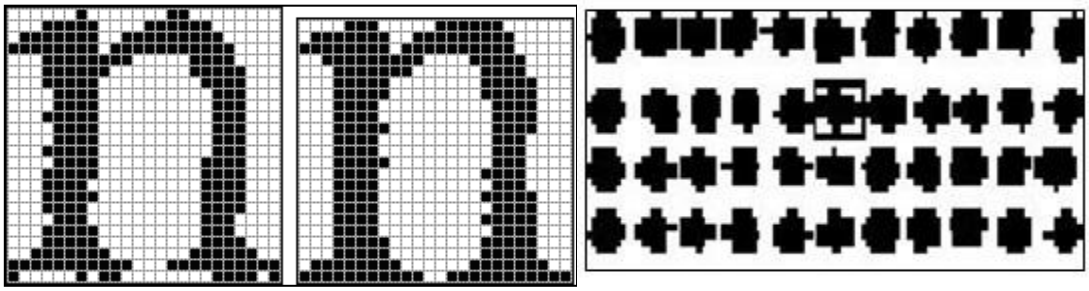

Рис. 1. Влияние шумов на изображения символов: а – искажения символа «n»; б – искажения символа «точка».

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.

Предлагаемый метод. 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.

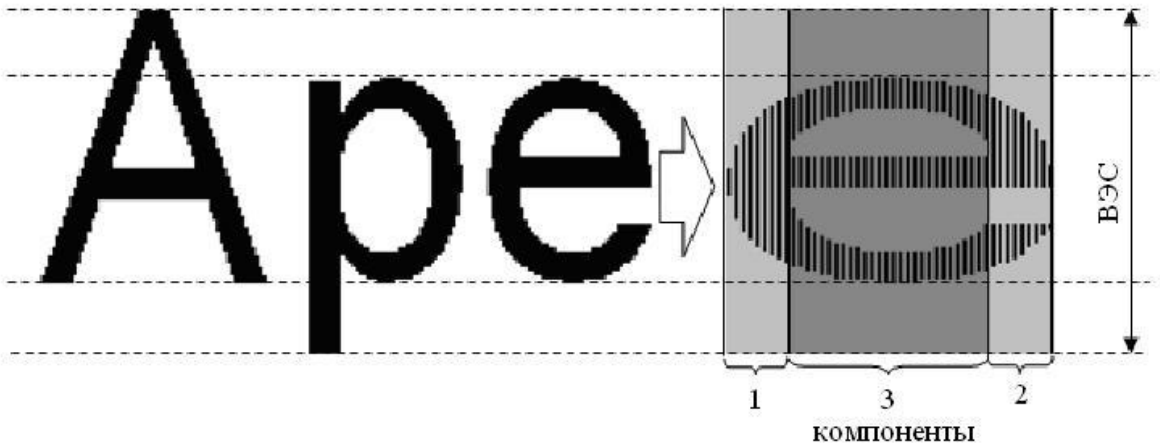

Рис. 2. Изображение буквы «е» и составляющие его вертикальные элементы строки с различным числом компонент

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.

После нахождения этих вероятностей легко получить правильную классификацию изображений символов, представив последние как упорядоченный набор вертикальных элементов.

Используя отдельный этап детализации связных символов изображения текста в виде вертикальных элементов строки и применив к ним нечеткую классификацию, была получена минимальная наиболее правдоподобная совокупность неискаженных элементов строки.

Выводы. Предложенный алгоритм представления и обработки изображения текста позволил получить достаточно высокую степень сжатия при хорошем качестве восстановленного 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 открывает новые возможности повышения информативности представления текстовых графических данных в инженерных реализациях.

СПИСОК ЛИТЕРАТУРЫ

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.