

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ СЕМЕНА КУЗНЕЦЯ

УДК 681.518.54



Тези доповідей

**VIII міжнародної науково-практичної
конференції
“Проблеми та перспективи розвитку
ІТ-індустрії”
28 – 29 квітня 2016 р.**

Харків 2016

УДК 681.518.54

Тези доповідей VIII міжнародної науково-практичної конференції “Проблеми та перспективи розвитку ІТ-індустрії”, 28 – 29 квітня 2016 р. – Х.: ХНЕУ імені Семена Кузнеця, 2016. – 114 с.

Наведені тези пленарних та секційних доповідей за теоретичними та практичними результатами наукових досліджень і розробок. Представлені результати теоретичних досліджень в галузях проектування інформаційних систем, технологій захисту інформації, використання сучасних інформаційних технологій в управлінні системами, моделювання бізнес-процесів, застосування геоінформаційних технологій, дистанційній освіті, інформаційних технологій в видавничо-поліграфічній галузі.

Матеріали публікуються в авторській редакції.

За достовірність викладених фактів, цитат та інших відомостей відповідальність несе автор.

АЛГОРИТМ НЕЧЕТКОЙ КЛАССИФИКАЦИИ ВЕРТИКАЛЬНЫХ ЭЛЕМЕНТОВ СТРОКИ ДЛЯ СЖАТИЯ ИЗОБРАЖЕНИЯ ТЕКСТА

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 деталями – 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 распространенных форматах DjVu и PDF. Степень сжатия информации с помощью методов классификации тем выше, чем меньше классов образуется при классификации и чем больше элементов в каждом классе. 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.

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.

Шумы печати и сканирования случайным образом искажают вертикальные элементы. Так что среди них могут быть искаженные и неискаженные элементы. Однако нет необходимости 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 упорядоченные наборы вертикальных элементов, разделенные пробелами.

Алгоритм состоит из трех частей. Первое, разделение изображения текста на вертикальные элементы строки. Второе, проведение нечеткой классификации их совокупности с определением минимальной наиболее правдоподобной совокупности неискаженных элементов. Третье, классификация изображений символов, использующая полученные во второй части вероятности, с которыми каждая пара вертикальных элементов является искажением одного и того же неискаженного вертикального элемента.

Предложенный алгоритм позволил получить достаточно высокую степень сжатия. Качество восстановленного изображения, лучше, чем оригинал (изображение сканированного текста), и близкое к электронному варианту.

Для наиболее часто используемого разрешения изображения текста 300 dpi были получены следующие сравнительные показатели сжатия:

- в работе [1] преимущество над JB2 – 8 %;
- в работе [2] преимущество над JB2 – 25 %;
- в работе [3] преимущество над JB2 – 37 %.

Это открывает новые возможности повышения информативности электронного представления графических текстовых данных в инженерных реализациях.

Список литературы

1. 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.
2. 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.
3. 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.