МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

МЕЖДУНАРОДНАЯ АКАДЕМИЯ ИНФОРМАТИЗАЦИИ

НАУЧНО-ТЕХНИЧЕСКОЕ ПРЕДПРИЯТИЕ «КРИПТОСОФТ»

ПЕНЗЕНСКИЙ ФИЛИАЛ ОАО «РОСТЕЛЕКОМ»

ООО «ЦЕНТР АНАЛИЗА И РАЗВИТИЯ КЛАСТЕРНЫХ СИСТЕМ»

ПЕНЗЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

НОВЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И СИСТЕМЫ

ТРУДЫ
X Международной научно-технической конференции

г. Пенза, 27–29 ноября 2012 г.

Proceedings of the Tenth International Conference of Science and Technology

NEW INFORMATION TECHNOLOGIES AND SYSTEMS

Penza, Russia, November 27–29, 2012

Пенза
Издательство ПГУ
2012

УДК 004.4
Н72

Новые информационные технологии и системы : тр. X Между-
Н72 нар. науч.-техн. конф. (г. Пенза, 27–29 ноября 2012 г.). – Пенза : Изд-во ПГУ, 2012. – 390 с.

ISBN 978-5-94170-517-7

Представлены материалы докладов, сделанных на X Международной научно-технической конференции «Новые информационные технологии и системы» («НИТиС-2012»), проводимой Министерством образования и науки РФ, Международной академией информатизации, Академией информатизации образования, научно-техническим предприятием «Криптософт» и Пензенским государственным университетом.

Доклады охватывают широкий спектр проблем в области новых информационных технологий в производстве, управлении, образовании; рассматриваются вопросы построения высокопроизводительных вычислительных комплексов, систем и сетей. Представлены современные технологии хранения и обработки данных, создания аппаратно-программных комплексов и информационно-вычислительных систем, интеллектуальных систем и систем управления, вопросы моделирования информационно-вычислительных систем и применения математических методов в информатике. В тематику конференции включен раздел, посвященный вопросам практической медицины.

УДК 004.4

ISBN 978-5-94170-517-7

В блоке анализа записи (блок 4) происходит проверка возможности перехода диагностической последовательности на следующий шаг. Если шаг сценария является последним, то он удаляется из списка, и информация сохраняется в отчет (блок 7). Вместе с тем анализируется время, содержащееся в записи. При превышении временного интервала времени ожидания элемента, он удаляется из списка с записью информации в отчет.

Если в процессе поиска активного сценария соответствующего фильтру так и не был найдено ни одного элемента списка, то данные передаются в блок анализа на вхождение сценария (блок 5). В данном блоке запись анализируется на возможность запуска нового сценария для данного фильтра. И если такая возможность существует, в список диагностических последовательностей будет добавлена новая запись, содержащая указатель на найденный сценарий, фильтр и условие перехода к новому шагу (блок 6).

После прохождения всех итераций в списке диагностических последовательностей останутся те элементы, выполнение которых было нарушено по каким-то причинам. Данные последовательности так же будут добавлены в отчет.

Предложенный алгоритм позволяет выполнять обработку данных в один проход. Учитывая объем анализируемой информации, который может доходить до сотен гигабайт, применение данного алгоритма позволяет существенно повысить скорость анализа.

Литература

1. Волчихин В. И., Пащенко Д. В., Пащенко В. Г. Средства воздушного наблюдения, контроля и управления. Надежность и качество, 2008.

2. Бутырин, О. (апрель 2007 г.). Методика количественной оценки качества техники пилотирования летчика в рейсовых условиях. Получено из Авивтранспортное обозрение. – URL: http://www.ato.ru/rus/media/ato/archives/78-2007/ak/ak3

3. Попов Э. В., Фоминых И. Б., Кисель Е. В., Шапот М. Д. Статические и динамические экспертные системы. – М. : "Финансы и статистика", 1996.

4. Попов Э. В. Экспертные системы решения неформализованных задач в диалоге с ЭВМ. – М. : Наука, 1987.

В. Г. Иванов, М. Г. Любарский, Ю. В. Ломоносов

Украина, Харьков, Национальный университет
«Юридическая академия Украины имени Ярослава Мудрого»

ЭФФЕКТИВНОЕ КОДИРОВАНИЕ ЭЛЕКТРОННОГО ИСТОЧНИКА ТЕКСТОВЫХ ДАННЫХ

Предложен и исследован метод сжатия изображений текста на основе новой классификации символов, благодаря которой удалось более чем в 2 раза уменьшить количество классов изображений символов и в среднем увеличить степень сжатия на 20 % по сравнению с лучшим на сегодняшний день алгоритмом сжатия изображений текста JB2.

Эффективное кодирование электронного источника текстовых данных на основе методов классификации является весьма перспективным и многообещающим направлением в теории и практике сжатия изображений [1–4]. Особое значение

данные методы могут иметь при сжатии изображений текста, которые повсеместно используются для перевода печатной продукции в электронный вид.

В данной работе предложен и исследован метод сжатия изображений текста, основанный на выделении неразделимых символов (букв и знаков пунктуации) и такой их классификации, что в каждый класс попадают только изображения одного и того же символа. Сложность этой задачи вызвана шумами, возникающими при печати текста на бумаге и последующем его сканировании. Необходимо также отметить, что в качестве неразделяемых символов могут выступать такие сочетания букв, как “fh”, а при невысоком разрешении сканирования – не только полные изображения букв, но и их фрагменты.

Предлагаемый метод можно условно разделить на несколько отдельно решаемых задач:

1) выделение из изображения текста неразделимых символов в виде минимальных прямоугольных областей, содержащих этот символ;

2) предварительная классификация полученных изображений символов по простым признакам (высота, ширина, полный периметр);

3) основная процедура – разбиение совокупности всех изображений неразделимых символов на классы, каждый из которых содержит изображения только одного символа и нахождение усредненного «представителя» для каждого класса;

4) создание «графического словаря», содержащего совокупность усредненных «представителей» и построение карты регионов, которая показывает размещение каждого символа из графического словаря на плоскости изображения текста.

Основная классификация, проводится с помощью алгоритма «просеивания» [5]. При сравнении двух изображений символов S_1 и S_2 с допустимыми отклонениями по высоте, ширине и периметру (ΔH, ΔW и ΔP) эти изображения накладываются друг на друга с помощью плоскопараллельного переноса так, чтобы их центры тяжести совпадали. Далее подсчитываются две величины: $R(S_1, S_2)$ – количество точек «существенных отличий», и $D(S_1, S_2)$ – количество общих точек совпадения, рис. 1.

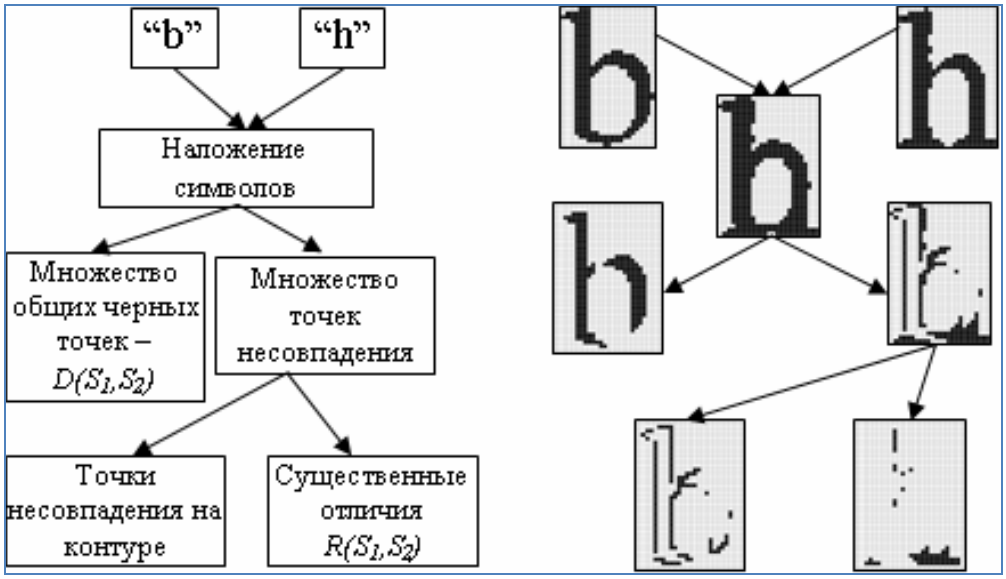

Рис. 1. Схема сравнения изображений символов “b” и “h”

Первая величина – это количество несовпадающих по яркости (белое – черное) точек, которые не являются смежными для совокупности общих черных точек. Таким образом, количество существенных отличий $R(S_1, S_2)$ игнорирует несовпадения в тех точках, которые лежат на периметрах изображений и, как правило, представляют собою шумы печати и сканирования. Вторая величина – нужна для обезразмеривания первой, чтобы диапазон возможных значений величины

$$\varepsilon(S_1, S_2) = \frac{R(S_1, S_2)}{D(S_1, S_2)} 100\ \% \qquad (1)$$

для всех пар символов не менялся при изменении размера шрифта и разрешения сканирования.

Функция $R(S_1, S_2)$, определяется с учетом веса. Весовой коэффициент каждой точки в $R(S_1, S_2)$ тем больше, чем больше у данной точки таких же смежных точек [6]. Таким образом, предлагаемая характеристика классификации ε (1), определяющая степень близости изображений двух символов при классификации алгоритмом «просеивания», мало чувствительна к шумам печати и сканирования. Она основана на стабильных характеристиках $R(S_1 S_2)$ и $D(S_1, S_2)$, которые подавляют (не учитывают) контурные шумы сравниваемых символов при их наложении с совмещенными центрами тяжести в результате плоско- параллельного переноса [7].

Сравнение с лучшим в настоящее время специальным алгоритмом для сжатия изображений текста – JB2, включенным в формат DjVu, показало, что качество классификации у предлагаемого метода значительно выше, чем у алгоритма JB2. Количество классов, получающееся в результате предложенной классификации, более чем в два раза меньше при всех разрешениях сканирования (табл. 1). Это является основной качественной характеристикой метода и дает широкие возможности повышения информативности этого алгоритма в инженерных реализациях [8].

Таблица 1

Сравнение эффективности алгоритмов

азрешение сканирования (dpi)	200	300	400	500	600
Исходный размер файла (kb)	505,3	1080,2	2003,9	3111,2	4498,0
Методы	**Размер файла после сжатия (kb)/ Коэффициент сжатия**				
Формат JPEG 2000	132,8 / 3,8	288,6 / 3,74	532,4 / 3,76	830,0 / 3,75	1200,3 / 3,75
Формат PDF	61,4 / 8,2	96,1 / 11,2	119,6 / 16,7	148,9 / 20,9	178,9 / 25,1
JB2 в формате DjVu	9,6 / 52,6	8,7 / 124,1	9,9 / 202,4	11,4 / 272,9	13,6 / 330,7
предлагаемый алгоритм	8,1 / 62,3	8,0 / 135,0	8,0 / 250,4	8,5 / 366,0	9,7 / 463,7

Предложенный алгоритм позволяет уменьшить размеры выходных данных, по сравнению с алгоритмом JB2 для всех разрешений сканирования (от 8 до 28,6 % при соответствующих значениях разрешений, табл. 2), что в среднем составляет около 20 %.

Таблица 2

Оценки эффективности предложенного алгоритма

Разрешение изображения текста (dpi)	Количество символов в исходном изображении	Количество классов после основной классификации $\varepsilon_{opt} = 6\,\%$	Количество классов после повторной классификации $\varepsilon_{opt} = 6\,\%$	Количество классов после классификации алгоритмом JB2
600 dpi	3558	197	72	314
500 dpi	3557	137	72	259
400 dpi	3557	130	71	199
300 dpi	3545	122	95	235
200 dpi	3890	237	148	451

Литература

1. Земсков В. Н. Сжатие изображений на основе автоматической классификации / В. Н. Земсков, И. С. Ким // Известия вузов. Электроника. – 2003. – № 2. – С. 50–56.

2. Gupta Maya R., Stroilov A. Segmenting for wavelet compression : Data Compression Conference, 2005. Proceedings. DCC 2005, 29–31 March 2005, USA, Utah, Snowbird. – 462 p. – URL: http://www.computer.org/portal/web/csdl/proceedings/ – 10.04.2010 г.

3. Иванов В. Г. Сокращение содержательной избыточности изображений на основе классификации объектов и фона / В. Г. Иванов, М. Г. Любарский, Ю. В. Ломоносов // Проблемы управления и информатики. – 2007. – № 3. – С. 93–102.

4. Иванов В. Г. Сжатие изображений на основе автоматической и нечеткой классификации фрагментов / В. Г. Иванов, Ю. В. Ломоносов, М. Г. Любарский // Проблемы управления и информатики. – 2009. – №1 – С. 52–63.

5. Прикладная статистика: Классификация и снижение размерности : справочник / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков и др.; под общ. ред. С. А. Айвазяна. – М. : Финансы и статистика, 1989. – 607 с.

6. Прэт У. К. Комбинированная система сжатия факсимильных данных с подбором символов / У. К.. Прэт, П. Дж. Капитан, Чжань Вэнсюнь, Э. Р. Хамилтон, Р. Х. Уоллис // Цифровое кодирование графики. Тематический выпуск. ТИИЭР. – 1980. – Т. 68, № 7. – С. 40–49.

7. Иванов В. Г. Сжатие изображения текста на основе выделения символов и их классификации / В. Г. Иванов, Ю. В. Ломоносов, М. Г. Любарский // Проблемы управления и информатики. – 2010. – № 6. – С. 111–122.

8. Иванов В. Г. Сжатие изображения текста на основе формирования и классификации вертикальных элементов строки в графическом словаре символьных данных / В. Г. Иванов, Ю. В. Ломоносов, М. Г. Любарский // Проблемы управления и информатики. – 2011. – № 5 – С. 98–109.

Б. В. Казаков

Россия, Пенза, Пензенский государственный университет

КВАЗИОПТИМИЗИРУЮЩИЙ АЛГОРИТМ РЕОРГАНИЗАЦИИ БАЗЫ ДАННЫХ

Описываются квазиоптимизирующий алгоритм реорганизации базы данных и инженерная методика реорганизации базы данных.

В современных СУБД поиск информации по запросам выполняется путем чтения из БД блоков (физических записей), при этом может выполняться чтение блоков с различными адресами. Характер информации может повлиять на частоту выборки того или иного блока БД. Это вызвано тем, что некоторые СУБД исполь-