УДК 004.627

Ю.В. ЛОМОНОСОВ, к.т.н., проф. НУ "ЮАУ им. Я. Мудрого", г. Харьков,
В.Г. ИВАНОВ, д.т.н., проф. НУ "ЮАУ им. Я. Мудрого", г. Харьков,
М.Г. ЛЮБАРСКИЙ, д.физ.-мат.н., проф. НУ "ЮАУ им. Я. Мудрого", г. Харьков,
Н.А. КОШЕВА, к.т.н., доц. НУ "ЮАУ им. Я. Мудрого", г. Харьков,
М.В. ГВОЗДЕНКО, ст. преп., НУ "ЮАУ им. Я. Мудрого", г. Харьков,
Н.И. МАЗНИЧЕНКО, ст. преп. НУ "ЮАУ им. Я. Мудрого", г. Харьков

ПОВЫШЕНИЕ ВЫЧИСЛИТЕЛЬНОЙ ЭФФЕКТИВНОСТИ ДВУХЭТАПНОГО АЛГОРИТМА СЖАТИЯ СИМВОЛЬНЫХ ДАННЫХ

Показано, что использование коротких первичных словарей в двухэтапном алгоритме сжатия символьных данных дает возможность уменьшить время кодирования на 20–25%. Представлены способы и критерии формирования первичных словарей символов, а также показатель их итерационного использования. Ил.: 5. Библиогр.: 10 назв.

Ключевые слова: сжатие символьных данных, словарь символов, двухэтапный алгоритм сжатия.

Постановка задачи. Применение методов классификации является перспективным и развивающимся направлением в теории и практике сжатия изображений [1 – 4]. Особое значение эти методы приобретают при сжатии изображений текста, которые используются для перевода печатных изданий в электронный вид. Известно, что из-за резких контрастных границ символов и их большого числа в строке неудовлетворительно работают стандартные методы сжатия, основанные на ортогональных преобразованиях, в том числе на преобразовании Фурье и вейвлет-анализе [4, 6].

В работах авторов [7 – 9] представлен метод сжатия изображений текста, основанный на выделении связных символов и их классификации. Установлено, что степень сжатия изображений текста является очень высокой при качестве восстановленного текста существенно лучшем (благодаря операциям усреднения), чем у исходного текста. Однако минимизация вычислительных затрат предлагаемых алгоритмов в этих работах не рассматривалась. Настоящей работой восполняется этот пробел.

Анализ литературы. В работах [7 – 9] использование двухэтапного алгоритма классификации символьных данных позволяет сформировать графический словарь изображений символов, который содержит практически минимально возможное число классов. Это позволило повысить степень сжатия изображений текста для всех разрешений по сравнению с алгоритмом JB2 (формат DjVu) почти на 20%.

В ранее предложенном двухэтапном алгоритме [7], на каждом этапе классификации применялись различные модификации метода "просеивания" [5]. Основная классификация (первый этап) проводилась непосредственно с помощью алгоритма просеивания и последующим усреднением представителей каждого класса. Это достаточно быстрая процедура. Однако, после первого этапа обработки, в сформированном графическом словаре встречались одинаковые символы, которые принадлежали различным классам. Применение повторной классификации, которая основана на алгоритме "наращивания областей", устраняет этот недостаток. Эта классификация требует большего времени на обработку, но в силу того, что классифицируются не все символы изображения текста, а только центры классов уже сформированного графического словаря, время обработки находится в допустимых пределах. Количество получаемых классов графического словаря уменьшается по сравнению с алгоритмом JB2 (формат DjVu) в 2,5 раза.

Основным недостатком двухэтапной классификации [7 – 9] является то, что на первом этапе классификации участвуют все символы, в том числе и те, которые образуют классы состоящие из одного представителя и являются уникальными. Это приводит к неоправданным временным затратам, когда подобный символ изображения текста сравнивается с остальными и в результате не находится ни одного подобного символа, образуя класс, состоящий из одного представителя.

На рис. 1 приведены примеры символов, которые являются одинаковыми, но не попали в один класс после первого этапа классификации по различным причинам. Это целое семейство символов "точка" и символа "г". В первом случае все символы при практически равных геометрических размерах (высота, ширина) значительно разнятся по периметру (отклонение, которого допускается не более 10%, что соответствует несовпадению всего двух точек в изображении данного символа). Во втором случае представленные символы не были классифицированы в один класс в ходе плоскопараллельного переноса и вычисления симметрической разности с совмещенными центрами тяжести при процедуре "просеивания". Наличие символов "*а*" и "г" – это результат слияния этих двух символов, которые в совокупности также образуют уникальных класс с одним представителем.

Поэтому возникает идея – на первом этапе классификации собрать в графический словарь сначала все символы, которые формируют классы с большим числом представителей, исключив их таким образом из дальнейшей классификации при формировании следующих классов. Когда дойдет очередь до классификации уникальных символов, то число сравниваемых с ними символов будет гораздо меньше, что позволит сократить общее время обработки всего символьного изображения.

Рис. 1. Примеры классов изображений символов с одним представителем.

Цель статьи. Создание общего графического словаря символов путем использования более коротких словарей, которые последовательно формируются на участках изображения текста. Разработка метода сокращения общего времени классификации синтезированных символьных изображений.

Описание методов. Необходимо напомнить, что классификация символов на первом этапе проводится методом "просеивания" [5, 10], который состоит в следующем. Выбирается произвольный элемент из классифицируемого множества и в один класс с ним помещаются все элементы близкие к нему. Далее рассматриваются только элементы, не вошедшие в первый класс. Из их числа произвольно выбирается какой-либо элемент и аналогичным образом строится второй класс. Этот процесс повторяется до тех пор, пока не будут исчерпаны все элементы исходного множества.

Второй этап классификации реализует алгоритм "наращивания областей", который заключается в том, что на первом шаге, начиная с

произвольно выбранного элемента классифицируемого множества, к его классу присоединяются все достаточно близкие элементы. На втором шаге к вновь присоединенным элементам добавляются все элементы, близкие к ним. Процесс "наращивания" повторяется до тех пор, пока на каком-то шаге не окажется новых элементов, которые можно было бы присоединить. Далее все элементы "выращенного" класса исключаются из классифицируемого множества и "выращивается" следующий класс. Алгоритм заканчивает работу, когда в классифицируемом множестве не остается ни одного элемента.

В представленной работе приводится иной подход к созданию общего словаря символов путем классификации символов изображения короткими словарями, которые последовательно формируются на участках изображения текста. Составление первичных словарей осуществляется на основе оценки их эффективности. Количество первичных словарей определяется такой условной характеристикой, как среднее число классифицированных символов первичного словаря.

Эффективность первичного словаря (K) представлена как отношение количества центров (классов) вошедших в словарь (N dic) к количеству символов на котором формировался данный первичный словарь (N symbols)

$$K = \frac{N\_dic}{N\_symbols}. \qquad (1)$$

На рис. 2 представлен график изменения эффективности словаря – K (выражение (1)), на всем множестве классифицируемых символов. На рис. 3 приведена пошаговая разность (приращение) эффективности первичного словаря – delta $K = K(i+1) - K(i)$ на том же множестве обрабатываемых символов.

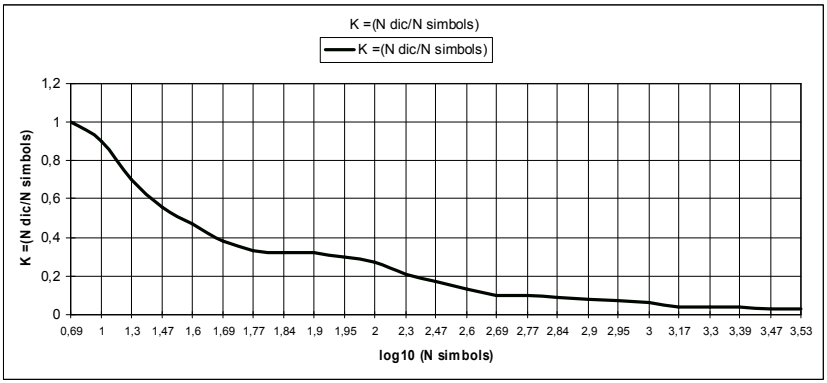

Рис. 2. Эффективность первичного словаря K

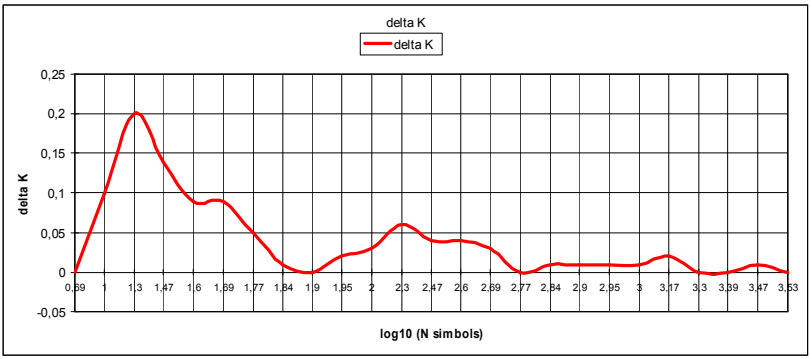

Рис. 3. Изменение эффективности первичного словаря

Максимум на рис. 3 определяет участок изображения текста, где сформированный первичный словарь будет наиболее эффективным. Дальнейшее увеличение области формирования первичного словаря (*N* symbols) в (1) не приводит к его интенсивному пополнению и его дальнейшее формирование необходимо остановить на полученном интервале. Далее осуществляется классификация символов всего изображения текста только центрами первичного словаря в соответствии с алгоритмом "просеивания".

Число итераций использования первичного словаря при обработке изображения текста определяется условной величиной – среднее количество классифицированных символов одним центром первичного словаря. В выражении (2), среднее количество символов в классе (*K*1) определяется как отношение количества классифицированных символов (Nclassific_symbols) к количеству центров первичного словаря (Nclasses)

$$K1 = \frac{Nclassific\_symbols}{Nclasses}. \qquad (2)$$

На рис. 4 представлено распределение символов изображения текста после их классификации центрами первичного словаря на две категории: классифицировано – непрерывная кривая; не классифицировано – пунктирная линия. Наглядно видно, что количество классифицированных символов (непрерывная кривая) быстро убывает, что свидетельствует о снижении эффективности классификации центрами первичного словаря. На оси абсцисс указано количество необработанных символов при *i*-ой итерации. На рис. 5 представлено среднее количество символов в классе на множестве необработанных символов – сплошная линия, а

приращение среднего количества символов в классе после классификации символов центрами первичного словаря – пунктирная кривая. Максимум приращения среднего числа символов в классе определяет число итераций.

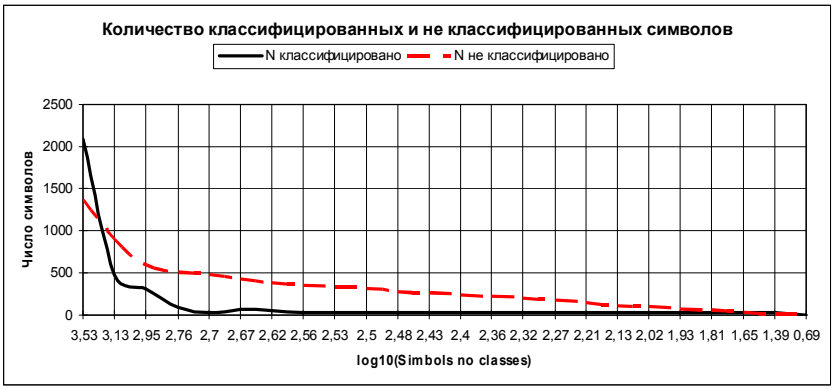

Рис. 4. Количество классифицированных и неклассифицированных символов

Рис. 5. Среднее число символов в классе и его приращение

Таким образом, на данном изображении классификация символов центрами первичных словарей наиболее эффективна при двух итерациях. Оставшееся множество символов можно классифицировать методом "просеивания" и далее на втором этапе методом "наращивания областей".

Выводы. Использование первичных словарей на первом этапе классификации методом "просеивания" (прямым перебором) позволило исключить из классифицируемого множества те символы, которые формируют классы с большим количеством представителей. Это позволило снизить общее время классификации на 20 – 25% по сравнению с последовательным применением метода "просеивания" и метода "наращивания областей" ко всему множеству изображений символов.

Список литературы. 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 [Электронный ресурс]: Data Compression Conference, 2005. Proceedings. DCC 2005, 29-31 March 2005, USA, Utah, Snowbird. – 462 p. – Режим доступа: 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.** *Гонсалес Р.* Цифровая обработка изображений / *Р. Гонсалес, Р. Вудс.* – М.: Техносфера, 2005. – 1072 с. **7.** *Иванов В.Г.* Сжатие изображения текста на основе выделения символов и их классификации / *В.Г. Иванов, М.Г. Любарский, Ю.В. Ломоносов* // Проблемы управления и информатики. – 2010. – № 6. – С. 74-84. **8.** *Иванов В.Г.* Сжатие символьных изображений на основе новой классифицирующей метрики / *В.Г. Иванов, М.Г. Любарский, Ю.В. Ломоносов, С.В. Деркач* // 17 міжнародна конференція з автоматичного управління "Автоматика -2010". Тези доповідей. – Том 2. – Харків: ХНУРЕ, 2010. – С. 162-164. 306 с. **9.** *Іванов В.Г.* Компресія зображень тексту на основі класифікуючої метрики з подавленням шумів друку та сканування / *В.Г. Іванов, М.Г. Любарський, Ю.В. Ломоносов, С.В. Котляр* // Праці 10-ї всеукраїнської міжнародної конференції "Оброблення сигналів і зображень та розпізнавання образів" (УкрОБРАЗ'2010). – К., 2010. – С. 161-165. **10.** *Шлезингер М.И.* Математические средства обработки изображений / *М.И. Шлезингер*. – К.: Наукова думка, 1983. – 200 с.

УДК 004.627

Підвищення обчислювальної ефективності двох-етапного алгоритму стиску символьних даних / Ломоносов Ю.В., Іванов В.Г., Любарський М.Г., Кошева Н.А., Гвозденко М.В., Мазниченко Н.І. // Вісник НТУ "ХПІ". Тематичний випуск: Інформатика і моделювання. – Харків: НТУ "ХПІ". – 2011. – № 36. – С. 107 – 114.

Показано, що при використанні у двох-етапному алгоритмі стиску символьних даних коротких первинних словників дає можливість зменшити час обробки на 20–25%. Представлені способи та критерії формування первинних словників класифікації символів, а також показник їх ітераційного використання. Іл.: 5. Бібліогр.: 10 назв.

Ключові слова: стиск символьних даних, словник символів, двох-етапний алгоритм стиску.

UDC 004.627

Rise of computing effectiveness double-step algorithm of compression of character data / Lomonosov U.V., Ivanov V.G., Lyubarsky M.G., Kosheva N.A., Gvozdenko M.V.,

Maznichenko N.I. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2011. – № 36. – P. 107 – 114.

It is shown, that at usage in double-step algorithm of compression of the character given short primary dictionaries the general processing time can be reduced by 20-25 %. Modes and measure of creation of primary dictionaries of classification of characters, and as an index of their iterative usage are presented. Figs.: 5. Refs.: 10 titles.

Keywords: cjmpression of character date, the dictionary of characters, double-step algorithm of compression.

Поступила в редакцию 15.06.2011