УДК 004:627
ББК 32.97

СОЗДАНИЕ ЭФФЕКТИВНОГО СЛОВАРЯ СИМВОЛОВ И СОКРАЩЕНИЕ ВРЕМЕННЫХ ЗАТРАТ ПРИ КЛАССИФИКАЦИИ ОЦИФРОВАННОГО ТЕКСТА

Иванов В. Г.,

Ломоносов Ю. В.,

Любарский М. Г.

Национальный юридический
университет имени Ярослава Мудрого,
Украина, г. Харьков

Анотація. *Показано, що використання коротких первинних словників в двоетапному алгоритмі стиснення символьних даних дає можливість зменшити час кодування на 20-25%. Представлені способи і критерії формування первинних словників символів, а так само показник їх ітераційного використання.*

Ключові слова: *зображення тексту, методи класифікації, словник символів.*

Методы классификации являются достаточно перспективными и активно используются в теории и практике сжатия изображений [1; 2; 4; 6; 8]. Наибольший интерес и значение эти методы приобретают при сжатии изображений текста (символьных изображений), которые используются при переводе печатной продукции в электронную форму.

Сам метод сжатия изображения текста на основе выделения символов и их классификации подробно изложен в работах авторов [3; 5; 7]. 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.

Метод оптимизации построения словаря символьных данных. Основным недостатком двухэтапной классификации [3; 5; 7] 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. Это целое семейство символов “точка” (слева на рис. 1) и символа “r” (справа на рис. 1). 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”.

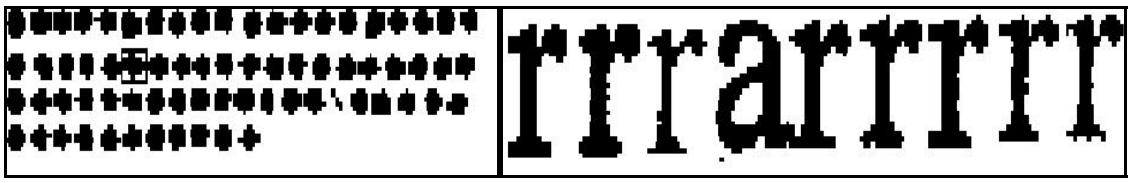

Рис. 1. Примеры классов изображений символов с одним представителем.

В данной работе предлагается следующее. 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.

Необходимо напомнить, что классификация символов на первом этапе проводится методом «просеивания» [3; 7], который имеет такой алгоритм. 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. Затем 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 данной работе представлен 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) оценивалась как отношение количества центров (классов) вошедших в словарь (N dic) к количеству (множеству) символов на котором формировался данный первичный словарь (N symbols), выражение (1)

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

Максимум отношения определяет участок изображения текста, где сформированный первичный словарь будет наиболее эффективным. Найденные центры используются для классификации на полном множестве символов. Количество итераций обработки изображения текста (выражение 2) определяется совпадением количества символов в классе с его приращением (см. рис. 2).

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

На рис. 2 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. 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».

Рис. 2. Среднее число символов в классе и его приращение.

Таким образом, можно сделать следующие 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 дало возможность снизить общее время классификации на 20–25 % по сравнению с последовательным применением метода “просеивания” и метода “наращивания областей” ко всему множеству изображений символов.

Список использованных источников

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. 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.
3. 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.
4. 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.
5. 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.
6. Прикладная статистика : классификация и снижение размерности : справ. изд. / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. – М. : Финансы и статистика, 1989. – 607 с.
7. Сжатие символьных изображений на основе новой классифицирующей метрики / В. Г. Иванов, М. Г. Любарский, Ю. В. Ломоносов, С. В. Деркач // Автоматика-2010 : 17 міжнар. конф. з автомат. упр. : тези доп. – Х., 2010. – Т. 2. – С. 162–164.
8. Gupta M. R. Segmenting for wavelet compression [Electronic resource] / Gupta M. R., Stroilov A. // Data Compression Conference, 2005. Proceedings. DCC 2005 : 29–31 March 2005, USA, Utah, Snowbird. – Way of access : http://www.computer.org/portal/web/csdl/proceedings/. – Title from the screen.

Аннотация. *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 так же показатель их итерационного использования.*

Ключевые слова*: изображение текста, методы классификации, словарь символов.*

Annotation. *It is shown, that use of short primary dictionaries in two steps algorithm of compression of symbolical data gives the chance to reduce time of coding for 20–25 %. Ways and criteria of formation of primary dictionaries of symbols, and as an indicator of their iterative use are presented.*

Key words: *the text image, classification methods, the dictionary of symbols.*