УДК 004.627

# Эффективность и сложность алгоритмов сжатия символьных данных

В.Г. Иванов[1], М.Г. Любарский[1], Ю.В.Ломоносов[1], Н.А. Кошевая[1], М.В. Гвозденко[1], Н.И.Мазниченко[1]

*Аннотация* – **The ways of approaching to the methods of processing of images of text on the basis of selection and classification of symbols are considered in this article.**

*Ключевые слова* – **словарь символов, эффективность словаря, классификация символов, карта размещения символов.**

## I. Вступление

Методы классификации являются перспективными и активно используются в теории и практике сжатия изображений [1 – 5]. Наибольший интерес и значение эти методы приобретают при сжатии изображений текста (символьных изображений), которые используются при переводе печатной продукции в электронную форму.

В данной работе оцениваются два варианта формирования словаря символов, который вместе с картой размещения символов, является основной величиной определяющей степень сжатия символьных изображений. Сам метод сжатия изображения текста на основе выделения символов и их классификации подробно изложен в работах авторов [6,7,8].

## II. Формирование словаря символов и порядок их классификации

В опубликованных ранее работах авторов [6-8] формирование общего словаря символов при их классификации осуществлялось путем их прямого перебора, а место расположения символов принадлежащих каждому классу указывалось в карте размещения символов.

В данной работе приводится новый 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 используемых первичных словарей определяется условной характеристикой. Это среднее число классифицированных символов центрами первичного словаря.

В настоящей работе эффективность первичного словаря (K) оценивалась как отношение количества центров (классов) вошедших в словарь (N dic) к количеству символов на котором формировался данный первичный словарь (N symbols), выражение (1)

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

На Рис.1 представлен график изменения эффективности словаря - K, на всем множестве классифицируемых символов. На Рис.2 приведена пошаговая разность (приращение) эффективности первичного словаря - delta K на том же множестве обрабатываемых символов.

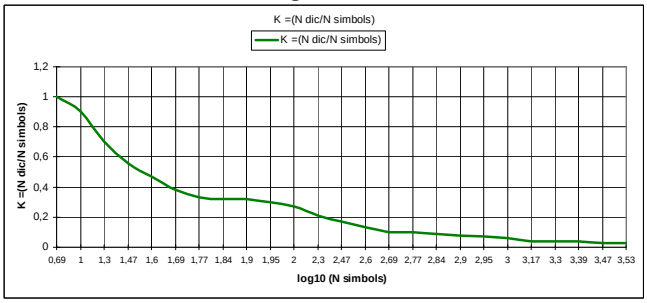


Рис.1. Эффективность первичного словаря K.


Рис.2. Изменение эффективности первичного словаря.

Максимум на Рис.2 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 изображения текста путем использования центров первичного словаря. Формируется карта размещения классифицированных символов. Процедура обработки повторяется на всем множестве символов, которые не были классифицированы первичным словарем. Количество итераций обработки изображения текста определяется условной величиной – среднее количество классифицированных символов центром первичного словаря. В выражении (2), среднее количество символов в классе (K1) определяется как

[1] Національний університет «Юридична академія України ім. Я. Мудрого», вул. Пушкінська, 77, Харків, 61024, УКРАЇНА, E-mail: nuau@uracad.kharkiv.edu

отношение количества классифицированных символов (Nclassific_symbols) к количеству центров первичного словаря (Nclasses)

$$K = \frac{N\_dic}{N\_symbols}. \quad (2)$$

На Рис.3 представлено распределение символов изображения текста после их классификации центрами первичного словаря на две категории: 1. классифицировано - непрерывная кривая; 2. не классифицировано – пунктирная линия. Наглядно видно, что количество классифицированных символов (непрерывная кривая) быстро убывает, что свидетельствует о снижении эффективности классификации символов изображения центрами первичного словаря. На оси абсцисс указано количество необработанных символов. На Рис.4 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 на множестве необработанных символов определяет число итераций.

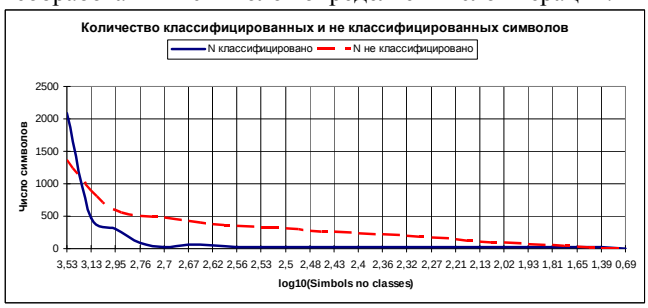


Рис. 3. Количество классифицированных и неклассифицированных символов.

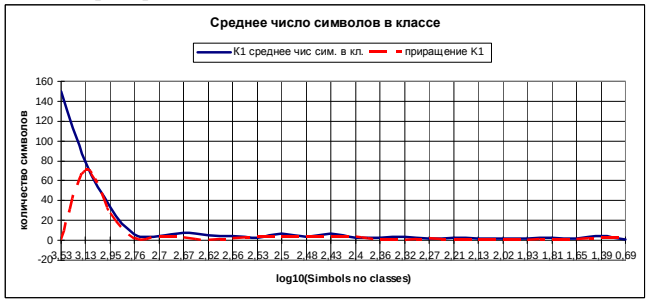


Рис. 4 Среднее число символов в классе и его приращение.

Таким образом, классификация символов центрами первичных словарей наиболее эффективна при двух итерациях, а оставшееся множество символов можно классифицировать методом прямого перебора. Это свидетельствует о том, что оставшиеся символы изображения редко встречаются и использование первичных (коротких) словарей для их классификации не будет столь эффективным.

## III. Выводы

Использование описаного способа формирования общего словаря символов на изображении текста позволяет снизить на 20-25% время обработки всего изображения по сравнению с методом прямого перебора символов при их классификации.

## Список литературы

[1] Земсков В.Н. Сжатие изображений на основе автоматической классификации [Текст] / 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 [Текст] / 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 [Текст] / 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 [Текст]: 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] 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. 74-84.

[7] Иванов В.Г. Сжатие символьных изображений на основе новой классифицирующей метрики. [Текст] / 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.

[8] Іванов В.Г. Компресія зображень тексту на основі класифікуючої метрики з подавленням шумів друку та сканування. [Текст] / 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) – Київ, 2010. – с.161-165.