УДК 004.627

В.Г. ИВАНОВ, д-р. техн. наук, проф., НУ "ЮАУ им. Я. Мудрого", Харьков,
Ю.В. ЛОМОНОСОВ, канд. техн. наук, доц., НУ "ЮАУ им. Я. Мудрого", Харьков,
М.Г. ЛЮБАРСКИЙ, д-р. ф.-м. наук, проф., НУ "ЮАУ им. Я. Мудрого", Харьков

КЛАССИФИКАЦИЯ СИМВОЛОВ В АЛГОРИТМАХ СЖАТИЯ ИЗОБРАЖЕНИЯ ТЕКСТА И СИСТЕМЫ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ

Показано, что получение минимально возможного количества классов при двухэтапной классификации изображений символов текста дает возможность уменьшить ошибку распознавания текста системами оптического распознавания практически на 50% по сравнению с форматом BMP и около 35% – 40% по сравнению с классификацией в формате DjVu. Ил.: 3. Табл.: 2. Библиогр.: 11 назв.

Ключевые слова: классификация, изображения символов, оптическое распознавание.

Постановка задачи. Использование методов классификации является весьма перспективным и развивающимся направлением в теории и практике сжатия изображений различной физической природы [1 – 4]. Наиболее весомое 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 неудовлетворительно работают классические методы сжатия, основанные на ортогональных преобразованиях, в том числе на преобразовании Фурье и вейвлет-анализе [4].

В работе авторов [5] представлен метод сжатия изображений текста, основанный на выделении связных символов и их классификации. Установлено, что практически минимальное количество классов, которые были получены в результате классификации выделенных символов, в значительной степени определяет высокий коэффициент сжатия всего изображения текста.

Было так же отмечено, что благодаря операциям усреднения предложенная классификация символов [5] существенно улучшает качество восстановленного текста в сравнении с исходным. Это обстоятельство обращает на себя внимание и вызывает научный интерес в исследовании влияния классификации связных символов на качество распознавания текста в системах оптического распознавания символов OCR (optical character recognition).

Анализ литературы. В работе [5] показано, что применение двухэтапного алгоритма классификации символьных данных позволяет сформировать графический словарь изображений символов, который содержит практически минимально возможное число классов. Сравнение с лучшим в настоящее время специальным алгоритмом для сжатия изображений текста – JB2, включенным в формат DjVu, показало, что качество классификации у предложенного метода значительно выше, чем у алгоритма JB2. Количество классов, полученных в результате классификации, более чем в 2 – 2,5 раза меньше при разрешениях сканирования в диапазоне 200 – 600 dpi, что, в свою очередь, позволило повысить степень сжатия всего изображения текста по сравнению с алгоритмом JB2 (формат DjVu) почти на 20%.

В данной работе для оценки качества распознавания восстановленного изображения текста использовалась наиболее распространенная в настоящее время система OCR, разработанная компанией ABBYY – FineReader 9.

Все современные оптические системы распознавания работают с довольно сложными алгоритмами обработки в различных сегментах изображения. Наиболее распространенные сегменты обрабатываемых изображений: область текста, картинка, таблица, штрих-код, разделитель и т.д. [6 – 8]. Общая ошибка распознавания состоит из ошибок различных типов, которые были получены при обработке всех сегментов изображения [8, 9]. В силу направленности исследований и для минимизации типов ошибок распознавания в качестве исходного изображения использовалась отсканированная страница журнала, содержащая только текст на английском языке. Таким образом, общую ошибку распознавания могут формировать только число неправильно распознанных символов и ошибки форматирования текста [6, 8].

Цель статьи. Получить и проанализировать ошибку распознавания восстановленного текста, полученного в результате использования двухэтапного алгоритма классификации символов. Провести сравнительный анализ количественной оценки качества распознавания текста.

Описание методов. Классификация символов на первом этапе проводится методом "просеивания" [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. Схематично классификация символов на первом этапе представлена на рис. 1.

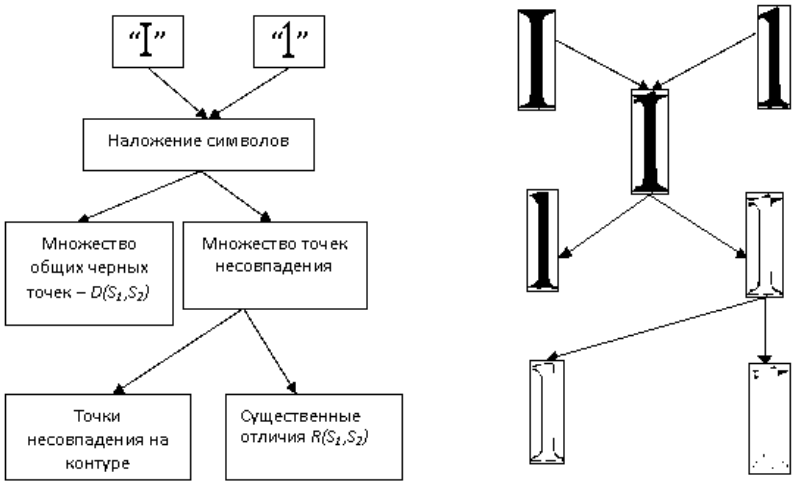


Рис. 1. Классификация символов на первом этапе.

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.

Первая величина – это количество несовпадающих по яркости (белый – черный) 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

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

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 [11]. Таким образом, полученная величина ε (1), определяющая степень близости изображений двух символов при классификации алгоритмом "просеивания", мало чувствительна к шумам печати и сканирования.

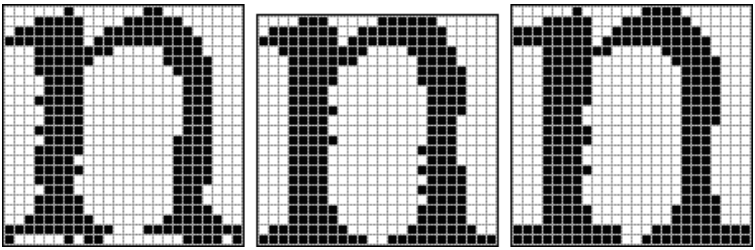
Второй этап классификации реализует алгоритм "наращивания областей" [10], который состоит в том, что на первом шаге, начиная с произвольно выбранного элемента классифицируемого множества, к его классу присоединяются все достаточно близкие элементы. На втором шаге к вновь присоединенным элементам добавляются 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. На рис. 2 представлены результаты всех этапов классификации символов.

Из приведенных результатов классификации символов на рис. 2 видно, что форма и контуры полученного изображения символа стали намного лучше по сравнению с исходными изображениями того же символа (рис. 2 а) и (рис. 2 г).

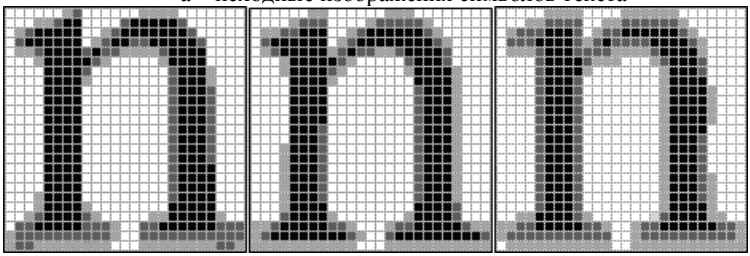
В таблице 1 показано количество классов после основной и повторной классификаций для различных разрешений. Для сравнения приведено количество классов после классификации алгоритмом JB2 (формат DjVu).

Данные, приведенные в таблице, демонстрируют достаточно высокую эффективность как первой, так и повторной классификаций и несомненное преимущество перед алгоритмом JB2.

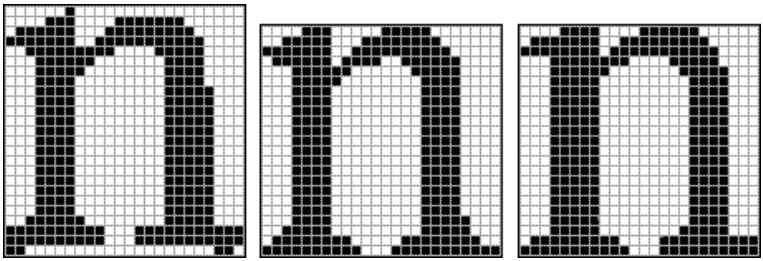
В таблице 2 приведены численные значения ошибок распознавания текста для исходного изображения в формате BMP, после двухэтапной классификации и после классификации алгоритмом JB2 в формате DjVu.



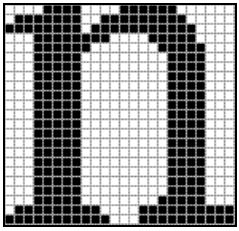
а – исходные изображения символов текста



б – совокупность совмещенных изображений символов в классах на первом этапе



в – усредненные представители классов. Результат первого этапа классификации



г – результат второго этапа классификации. Конечный представитель класса

Рис. 2. Изображения символов двухэтапной классификации

Таблица 1

Количество классов после основной и повторной классификаций

Разрешение изображения текста (dpi)	Количество классов в исходном изображении	Количество классов после основной классификации	Количество классов после второй классификации	Количество классов после классифика-ции 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

Таблица 2

Количество ошибок распознавания после различных классификаций

Разреше-ние изображе-ния текста (dpi)	Количество классов в исходном изображении	Количество ошибок распознавания в формате BMP /(%)	Количество ошибок распознавания после двухэтапной классификации /(%)	Количество ошибок распознавания после классификации алгоритмом JB2 /(%)
600 dpi	3558	0 / 0%	0 / 0%	0 / 0%
500 dpi	3557	0 / 0%	0 / 0%	0 / 0%
400 dpi	3557	6 / 0,168%	0 / 0%	4 / 0,112%
300 dpi	3545	16 / 0,451%	8 / 0,225%	14 / 0,394%
200 dpi	3890	42 / 1,079%	26 / 0,668%	39 / 1,0%

На рис. 3 приведена ошибка распознавания для рассматриваемых форматов изображения текста. Подобная количественная оценка качества распознавания (ошибка ~ 1%) свидетельствует о достаточно высоком качестве исходного изображения текста, так как известно, что точность распознавания латинских символов в сканированных печатных документах практически для всех систем OCR превышает 99%.

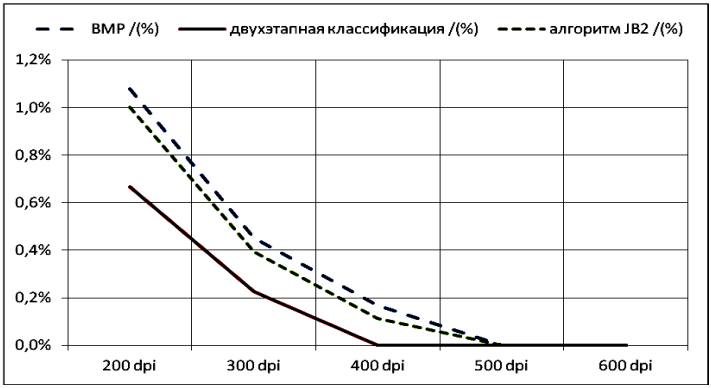


Рис.3. Ошибка распознавания изображения текста

Выводы. Формирование минимально возможного числа классов при двухэтапной классификации изображений символов позволяет: 1) повысить степень сжатия изображения текста по сравнению с алгоритмом JB2 (DjVu) почти на 20% [5]; 2) уменьшить ошибку распознавания текста системами OCR по сравнению с форматами *.bmp и *.djvu на 50% и 35% – 40% соответственно.

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.** *Гонсалес, Р.* Цифровая обработка изображений / *Р. Гонсалес, Р. Вудс.* – М.: Техносфера, 2005. – 1072 с. **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.** *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. **6.** *Арлазаров В.Л.* Распознавание строк печатных текстов / *В.Л. Арлазаров, П.А. Куратов, О.А. Славин* // Сб. трудов ИСА РАН "Методы и средства работы с документами". – М.: Эдиториал УРСС. – 2000. – С. 31-51. **7.** *Форсайт Дэвид А.* Компьютерное зрение. Современный подход / *Форсайт Дэвид А., Понс Джин*. Computer Vision: A Modern Approach. – М.: Вильямс, 2004. – 928 с. **8.** *Горелик А.Л.* Методы распознавания / *А.Л. Горелик, В.А. Скрипкин* – 4-е изд. – М.: Высшая школа, 2004. – 262 с. **9.** *Вапник В.Н.* Теория распознавания образов / *В.Н. Вапник, А.Я. Червоненкис*. – М.: Наука, 1974. — 416 с. **10.** 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. **11.** *Шлезингер М.И.* Математические средства обработки изображений / *М.И. Шлезингер*. – К.: Наукова думка, 1983. – 200 с.

УДК 004.627

Класифікація символів в алгоритмах стиску зображень тексту та системи оптичного розпізнавання / Іванов В.Г., Ломоносов Ю.В., Любарський М.Г. .// Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2012. – № 62 (968). – С. 83 – 90.

Показано, що отримання мінімально можливої кількості класів при двоетапній класифікації зображень символів тексту дає можливість зменшити похибку розпізнавання тексту системами оптичного розпізнавання практично на 50% в порівнянні з форматом BMP і близько 35% – 40% в порівнянні з класифікацією у форматі DjVu. Іл.: 3. Табл.: 2. Бібліогр.: 11 назв.

Ключові слова: класифікація, зображення символів, оптичне розпізнавання.

UDC 004.627

Classification of characters is in the algorithms of compression of text and system of optical recognition / Ivanov V.G., Lomonosov U.V., Lyubarsky M.G. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2012. – №. 62 (968). – P. 83 – 90.

It is proved that the forming of the possible minimum of the quantity of classes within two-stage classification of the text symbol images allows to minimize the error of text recognition by the systems of optical recognition for about 50% as comparing with the BMP format and about 35% – 40% as comparing with classification within DjVu format. Figs.: 3. Tabl.: 2. Refs.: 11 titles.

Keywords: classification, images of characters, optical recognition.

Поступила в редакцию 25.07.2012