УДК 004:627

ПРОБЛЕМЫ ЦИФРОВОГО ПРЕОБРАЗОВАНИЯ И АВТОМАТИЧЕСКОГО ПЕРЕВОДА ТЕКСТА

кандидат технических наук, доцент, Ломоносов Ю. В.
Национальный юридический университет имени Ярослава Мудрого, Украина, Харьков
Ломоносова Ж. В.
Харьковский национальный университет строительства и архитектуры, Украина, Харьков

В работе рассматриваются методы классификации, применяемые при цифровой обработке би-тональных изображений текста, полученных сканированием или цифровым фотографированием. Для известных на сегодняшний день алгоритмов классификации, включая хорошо зарекомендовавший себя - алгоритм JB2, приведены количественные характеристики классификации – число классов, получаемых этими алгоритмами для изображения стандартной страницы текста. Так-же показано, что минимально возможное количество классов при классификации изображений символов текста позволяет уменьшить ошибку распознавания текста системами оптического распознавания. Минимизация ошибки распознавания текста позволяет повысить качество автоматического перевода на иностранный язык.

Ключевые слова: классификация изображений текста, системы оптического распознавания, компьютерный перевод текста.

кандидат технічних наук, доцент, Ломоносов Ю. В. Проблеми цифрового перетворення та автоматичного перекладу тексту / Національний юридичний університет імені Ярослава Мудрого, Україна, Харків.

Ломоносова Ж. В. Проблеми цифрового перетворення та автоматичного перекладу тексту / Харківський національний університет будівництва та архітектури, Україна, Харків

В роботі розглядаються методи класифікації, що застосовуються при цифровій обробці бі-тональних зображень тексту, отриманих скануванням або цифровим фотографуванням. Для відомих на сьогоднішній день алгоритмів класифікації, включаючи добре відомого - алгоритм JB2, наведені кількісні характеристики класифікації - число класів, одержаних цими алгоритмами для зображення стандартної сторінки тексту. Також показано, що мінімально можлива кількість класів при класифікації зображень символів тексту дозволяє зменшити помилку розпізнавання тексту системами оптичного розпізнавання. Мінімізація помилки розпізнавання тексту дозволяє підвищити якість автоматичного перекладу на іноземну мову.

Ключові слова: класифікація зображень тексту, системи оптичного розпізнавання, комп'ютерний переклад тексту.

PhD, Associate Professor, Lomonosov Yu. V. Problems of digital conversion and automatic translation of text / Yaroslav Mudryi National Law University, Ukraine, Kharkiv

Lomonosova Zh. V. Problems of digital conversion and automatic translation of text / Kharkiv National University of Civil Engineering and Architecture, Ukraine, Kharkiv

The paper deals with the classification methods used in digital processing of bi-tonal image of the text, obtained by scanning or digital photography. For the currently known classification algorithms, including well-established - an algorithm JB2, given the quantitative characteristics of the classification - the number of classes obtained by these algorithms to image the standard page of text. Thus, it was shown that the smallest possible number of classes in the classification of images of text characters can reduce the recognition error text optical character recognition systems. Minimizing OCR errors can improve the quality of automatic translation into a foreign language.

Key words: classification of images of text, optical character recognition (OCR), computer translation of the text.

Вступление. **Постановка проблемы и анализ литературы.** Бесценные сокровища литературной, научной, философской мысли, которые накопило человечество, хранятся в многочисленных библиотеках в печатном или рукописном виде и нуждаются в преобразовании в электронную форму. Процесс перевода бумажных книг в электронный (цифровой) вид *называется оцифровкой*. В результате оцифровки получаются *электронные книги* – то есть хранимый в файле текст, оформленный в виде привычной книги.

Еще в недавнем прошлом создание электронной книги происходило только с помощью ручного набора текста, что является крайне трудоемкой и, следовательно, дорогой операцией. В настоящее время оцифровка печатных документов осуществляется с помощью сканера или цифрового фотоаппарата с последующей программной обработкой и сохранением в одном из форматов графических файлов. На следующем этапе производится оптическое распознавание текста (технология OCR), превращающая изображение текста в собственно текст, для дальнейшего сохранения в одном из текстовых форматов или для автоматического перевода на иностранный язык.

Использование методов классификации является весьма перспективным и развивающимся направлением в теории и практике обработки изображений различной физической природы [1-3]. Наиболее весомое значение эти методы приобретают при обработке и 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 компрессии данных, основанные на ортогональных

преобразованиях, в том числе на преобразовании Фурье и вейвлет-анализе [4].

В работах [5, 6] представлены методы сжатия изображений текста, основанные на выделении связных символов и их классификации. Установлено, что практически минимальное количество классов, которые были получены в результате классификации выделенных символов, определяет высокий коэффициент сжатия всего изображения текста. Было так же отмечено, что благодаря операциям усреднения классификация символов существенно улучшает качество распознавания текста в системах оптического распознавания символов OCR (optical character recognition).

В работах [5, 6] показано, что сравнение с лучшим в настоящее время специальным алгоритмом для сжатия изображений текста – JB2, включенным в формат DjVu, качество классификации у рассмотренных методов значительно выше. Количество классов, полученных в результате классификации, более чем в 2-2,5 раза меньше при разрешениях сканирования в диапазоне 200-600 dpi. Это позволило повысить степень сжатия всего изображения текста по сравнению с алгоритмом JB2 (формат DjVu) на 25% – 35%. Уменьшение количества классов в словаре символов текста позволяет также снизить ошибку распознавания символов текса, что в свою очередь приведет к повышению качества автоматического перевода.

Из сказанного можно сделать вывод, что массовая оцифровка печатной продукции прошлых лет в электронную форму с последующим переводом на иностранный язык – это слишком сложный путь, по крайней мере, пока программы оптического распознавания и автоматического перевода не будут существенно усовершенствованы. Таким образом, на сегодняшний день, остается единственный путь: 1) улучшение качества классификации символов в изображении текста; 2) минимизация ошибок распознавания символов текста; 3) улучшение качества автоматического перевода текста при помощи компьютерных программ.

В этом направлении сделаны существенные шаги, начиная от уже показавших свою практическую ценность форматов PDF и DjVu и заканчивая алгоритмами [3, 5, 6], находящимися еще в стадии разработки.

Цель настоящей статьи – дать обзор идей и методов, на которых основаны алгоритмы преобразования и классификации изображений текста. Провести сравнение полученных количественных показателей методов классификации. Показать влияние качества классификации символов на точность их распознавания и качество автоматического перевода текста.

Общая схема и последовательность обработки текста представлена на рис.1.

Выделение символов и их классификация. Высокие результаты, демонстрируемые алгоритмом JB2, объясняются тем, что он использует классификацию символов. Вообще идея сжатия информации с помощью классификации очень проста и идеально подходит для обработки изображений текста.

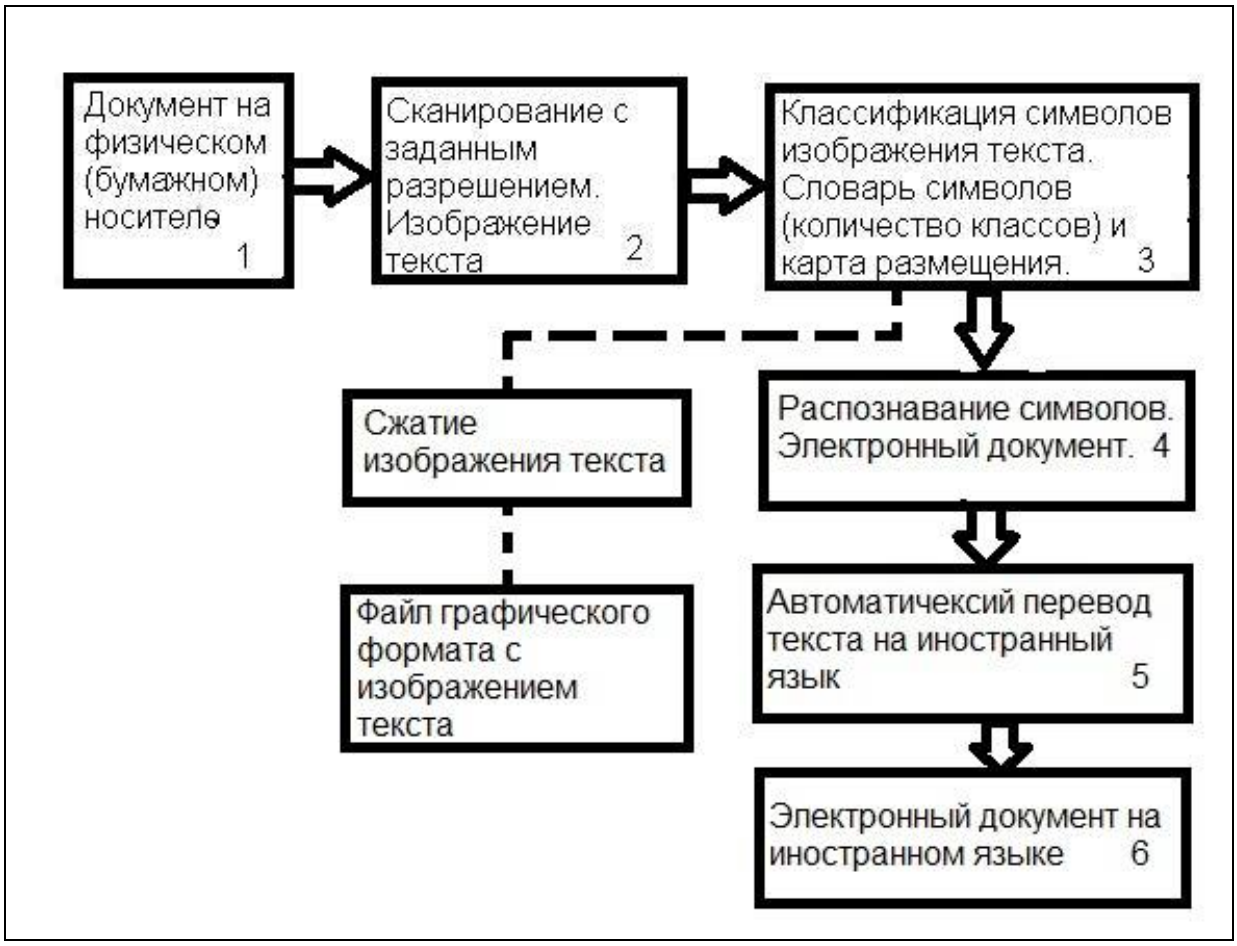

Рис. 1. Этапы преобразования и перевода текста.

Пусть необходимо обработать некую информацию, которую можно разбить каким-то образом на элементы. Если эти элементы информации объединить в классы так, чтобы в каждом классе находились тождественные (pattern matching) или почти тождественные (soft pattern matching) элементы, то нет нужды хранить все элементы информации – достаточно хранить только по одному элементу из каждого класса. Совокупность таких элементов – представителей классов – называется *словарем*. Кроме того для восстановления информации нужно еще иметь таблицу, называемую «картой размещения классов», которая для каждого класса указывает, в каком месте исходной информации находятся его элементы.

Ясно, что степень сжатия данных с помощью классификации тем выше, чем меньше классов образуется при классификации символов и чем больше элементов находится в каждом классе.

В случае сжатия изображения бинарного (черно-белого) текста естественным элементом информации является изображение отдельного символа (буквы, цифры, знака препинания и т.п.). Выделение символов не представляет собой особо трудную задачу. Во всех известных алгоритмах, включая алгоритм JB2, символы выделяются как связные области, состоящие из черных точек.

Следует заметить, что при этом некоторые грамматические символы распадаются на части (например, буква «ё» дает три символа), а некоторые (например, сочетания вида “fh”) объединяются в один. Кроме того метод непригоден для текстов с псевдо рукописным шрифтом. Степень сжатия и дальнейшая обработка таких текстов алгоритмом JB2 и другими катастрофически низкая.

Однако не это представляет собой главную трудность при классификации уже разделенных символов.

На рис. 2, взятом из работы [6], представлены три случайно выбранные изображения буквы «n» из различных 257, входящих в изображение страницы текста формата А4, при разрешении сканирования 300 dpi.

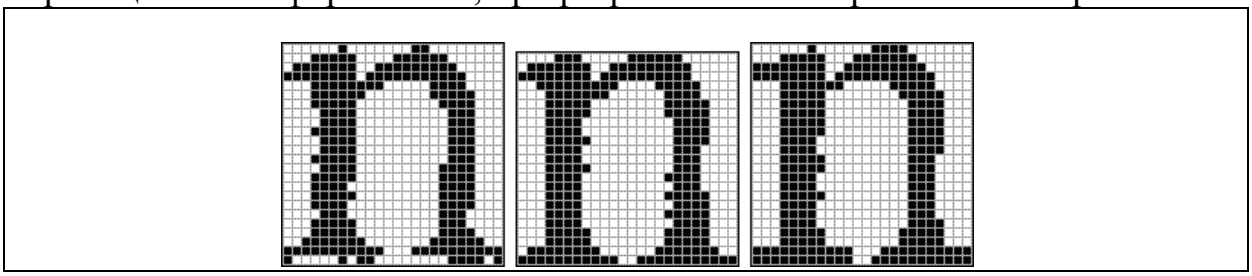

Рис. 2. Влияние шумов печати и сканирования на изображения символа “n”.

Всем очевидно, что на странице не найдется ни одной пары символов «n», полностью совпадающих друг с другом. То же, за редким исключением, относится и к другим символам, даже точкам. Причиной этого явления являются шумы (то есть случайные искажения), возникающие при печати страницы и ее последующем сканировании. Шумы печати в основном вызваны диффузией краски, жидкой или твердой, вдоль хаотически расположенных капилляров бумаги. Шумы сканирования – несовпадением контуров символа с матрицей сканера, подобно тому, как прямая наклонная линия на экране монитора отображается «ступеньками».

Человеку легко заметить, что все три изображения, приведенные на рис. 2, представляют собой букву «n». Однако пока не существует алгоритма, который мог бы установить тождественность этих символов с той же достоверностью, что и человек. Это и есть главная трудность, не позволяющая разбить изображения символов на классы, так чтобы одновременно выполнялись два условия:

Условие 1. В каждом классе находятся изображения только одного и того же символа;

Условие 2. Все изображения какого-либо символа находятся в одном классе.

Все алгоритмы классификации являются тем или иным компромиссом между этими условиями, причем условие 1 должно выполняться достаточно жестко, иначе в восстановленном тексте будут перепутаны символы, чем иногда грешит алгоритм JB2. Например, иногда путает между собой буквы «b» и «h», рис. 3.

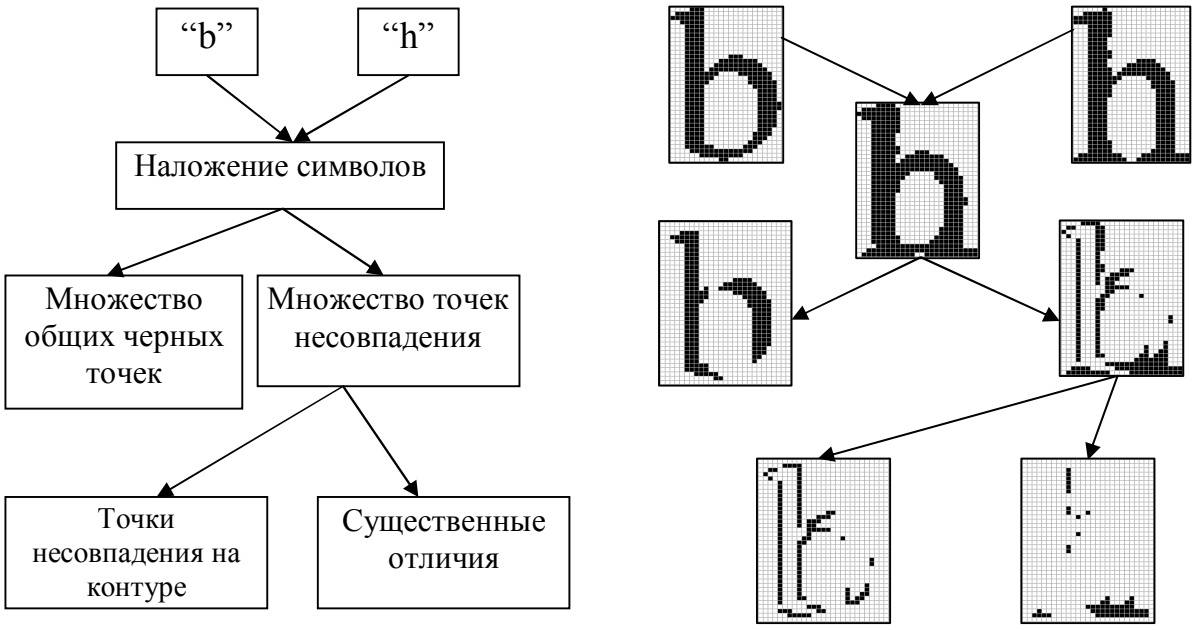

Рис. 3. Схема сравнения изображений символов “b” и “h”.

Соблюдение условия 1 влечет за собой ужесточение алгоритма сравнения изображений символов, так что условие 2, практически, невыполнимо. Это приводит к появлению значительно большего числа классов, чем количество символов, изображенных на странице, так как практически все символы дают по нескольку классов своих изображений. Чем больше при классификации образуется классов, тем больше словарь и (логарифмически) больше карта расположения классов. Как следствие, понижается степень компрессии. И хотя алгоритм JB2 и другие используют те или иные методы дополнительного сжатия словаря и карты, эффективность алгоритма в целом определяется *качеством классификации*, то есть количеством получившихся классов, которое в идеале (условие 2) должно совпадать с количеством символов, присутствующих в тексте, чье изображение подвергается обработке.

Таблица 1.

Количество классов при рассматриваемых методах классификации

Разрешение сканирования (dpi)	Количество классов в исходном изображении	Количество классов после предлагаемой классификации	Количество классов после классификации алгоритмом JB2
600	3558	72	314
500	3557	72	259
400	3557	71	199
300	3545	95	235
200	3890	148	451

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.

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, приблизительно равна 99%.

Отсутствие ошибок распознавания с высоким разрешением изображения текста (400 - 600 dpi) обусловлено высоким качеством “скелетизации” изображений символов, что обеспечивает расматриваемый дианазон разрешения.

Таблица 2.

Количество ошибок распознавания символов текста

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

При разрешении 200 – 300 dpi ошибка распознавания начинает возростать. И хотя ошибка не превышает 1% от общего количества символов на странице (3545 символов - для разрешения 300 dpi) необходимо помнить, что каждый не правильно распознанный символ текста искажает слово в состав котрого этот символ входит. Таким образом, можно предположить, что число ошибок распознавания символов в тексте определяет количество слов с грамматическими ошибками.

Для разрешения 300 dpi, в предложенных алгоритмах [3, 5, 6], число не правильно распознанных символов равно 8 (алгоритм JB2 формата DjVu ошибся 14 раз на тех же 3545 символах). Следовательно, без предварительного анализа или словарной проверки распознанной страницы текста, перед процедурой автоматического перевода на иностранный язык исходный текст имеет 8 или 14 слов с ошибками соответственно. Если совсем отказаться от классификации изображения текста (с разрешением 300 dpi) и использовать для распознавания графический формат *.bmp, то число ошибок возрастёт до 16 (Таблица 2).

Возможности и проблемы автоматического перевода текста. На последнем этапе обработки (рис. 1) распознанный текст необходимо

перевести на иностранный язык при помощи программ компьютерного перевода. Известно, что в настоящий период современные программы автоматического (машинного) перевода очень далеки от того идеала, к которому стремятся их разработчики. Системы нового поколения позволяют "запоминать" уже переведенные однажды конструкции и использовать их впоследствии, существует возможность создавать "пользовательский" словарь, что существенно облегчает перевод по определенной тематике и т.д..

Однако, все же недостатков у систем автоматического перевода больше чем достоинств. Нет смысла их подробно обсуждать – они хорошо известны специалистам по филологии, лингвистики и иностранным языкам [7]. К основным недостаткам можно отнести: 1) Программа автоматического перевода может не учитывать элементарных значений слов и не предлагать их в качестве варианта при переводе. 2) Не полностью учтены грамматические особенности входных и выходных языков. 3) Система автоматического перевода проводит синтаксический анализ на входе, выполняет пословный перевод, зачастую не обращая внимания на синтаксические связи.

Исходя из многообразия определяющих факторов и сложности автоматического перевода, довольно сложно получить точную количественную оценку ошибки переведенного текста. Таким образом, автоматический компьютерный перевод в принципе возможен, но его стоит рассматривать только как "черновой" вариант перевода, который подлежит обязательному редактированию. Пока только человек способен передать точно смысл иностранного теста, его стилистические оттенки и нюансы.

Выводы. Исходя из сложности решаемой комплексной задачи преобразования текста на физическом (бумажном) носителе в електронную форму с автоматическим переводом текста на иностранный язык, можно утверждать, что полностью автоматизировать весь процесс обработки, на сегодняшний день, не возможно. Однако, минизировать ошибки распознавания связных символов, возможно при повышении качества их классификации, то есть при полном выполнении условий 1 и 2. Это позволит сохранить качество автоматического перевода текста на иностранный язык, исключив из этого процесса ошибки вносимые при распознавании символов.

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. 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.

3. 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.
4. Иванов В. Г. Фурье и вейвлет анализ изображений в плоскости JPEG-технологий / В. Г. Иванов, Ю. В. Ломоносов, М. Г. Любарский // Проблемы управления и информатики. – 2004. – № 5. – С. 111-124.
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. Иванов В. Г. Классификационные методы сжатия изображений оцифрованного текста. Часть I / В. Г. Иванов, Ю. В. Ломоносов, М. Г. Любарский // Системи обробки інформації. - 2013. - № 2. - С. 36-43.
7. Сергеева Т. В. Решение грамматических задач в технике смысловой интерпретации, моделирующей речевое мышление человека / Т. В. Сергеева // Вестник ХГУ. Серия «Психология». – 1999. - № 432. – С. 297–301.

References:

1. Zemskov V. N. Szhatie izobrazheniy na osnove avtomaticheskoy klassifikatsii / V. N. Zemskov, I. S. Kim // Izvestiya vuzov. Elektronika. – 2003. – № 2. – S. 50-56.
2. Ivanov V. G. Sokrashchenie soderzhatelnoy izbytochnosti izobrazheniy na osnove klassifikatsii obektov i fona / V. G. Ivanov, M. G. Lyubarskiy, Yu. V. Lomonosov // Problemy upravleniya i informatiki. – 2007. – № 3. – S. 93-102.
3. Ivanov V. G. Szhatie izobrazheniy na osnove avtomaticheskoy i nechetkoy klassifikatsii fragmentov / V. G. Ivanov, Yu. V. Lomonosov, M. G. Lyubarskiy // Problemy upravleniya i informatiki. – 2009. – №1 – S. 52-63.
4. Ivanov V. G. Fure i veyvlet analiz izobrazheniy v ploskosti JPEG-tekhnologiy / V. G. Ivanov, Yu. V. Lomonosov, M. G. Lyubarskiy // Problemy upravleniya i informatiki. – 2004. – № 5. – S. 111-124.
5. Ivanov V. G. Szhatie izobrazheniya teksta na osnove vydeleniya simvolov i ikh klassifikatsii / V. G. Ivanov, M. G. Lyubarskiy, Yu. V. Lomonosov // Problemy upravleniya i informatiki. – 2010. – № 6. – S. 74-84.
6. Ivanov V. G. Klassifikatsionnye metody szhatiya izobrazheniy otsifrovannogo teksta. Chast I / V. G. Ivanov, Yu. V. Lomonosov, M. G. Lyubarskiy // Sistemi obrobki informatsiї. - 2013. - № 2. - S. 36-43.
7. Sergeeva T. V. Reshenie grammaticheskikh zadach v tekhnike smyslovoy interpretatsii, modeliruyushchey rechevoe myshlenie cheloveka / T. V. Sergeeva// Vestnik KhGU. Seriya «Psikhologiya». – 1999. - № 432. – S. 297–301.