Автоматизированная система распознавания рукописных исторических документов

© А.А. Рогов, А.Н. Талбонен, А.Г. Варфоломеев

Петрозаводский государственный университет

rogov@psu.karelia.ru, perhetal@onego.ru, avarf@psu.karelia.ru

Аннотация

Статья посвящена вопросам распознавания рукописных исторических документов, включая стенографические записи. Описаны устройство предполагаемой системы автоматизированной дешифровки рукописных источников, а также основные процессы, связанные с набором рукописного текста.

1 Введение

Широкое распространение и увеличение доступности технологий сканирования и цифрового фотографирования привело к быстрому росту цифровых коллекций исторических документов. В таких коллекциях документы хранятся в виде растровых графических файлов [4, 9]. Оцифровка решает множество проблем, связанных с сохранением культурного наследия и организацией доступа к нему для исследователей и массового пользователя. Однако для реализации полнотекстового поиска, изучения структуры и содержания документов, подготовки научных публикаций исторических источников требуется перевод источника из графического формата в текстовый, то есть распознавание текста.

Алгоритмы и программы автоматического распознавания текста разрабатываются уже несколько десятилетий. Общепризнанно, что распознавание текста включает в себя этапы предобработки (бинаризации изображения), сегментации (выделения текстовых областей, строк, слов, символов), анализа бинарных изображений символов или слов (установления значений признаков, сравнения с эталонами) и выбора наиболее подходящих словоформ из словаря в соответствии с определенной моделью языка. Можно сказать, что задача распознавания текстов на европейских языках, напечатанных на лазерных принтерах с использованием наиболее употребительных шрифтов и отсканированных на планшетных сканерах, практически решена. Однако все не так просто даже для книг середины XX века – изображения могут требовать существенной предобработки, выходящей за рамки функций, реализованных в популярных OCR-программах, шрифт может оказаться необычным, а язык – устаревшим с позиции современного словаря. Дополнительные сложности создают искривления строк, перепады яркости, просвечивания текста с обратной стороны и другие дефекты оригинала и изображения. Показательно, что один из самых заметных отечественных проектов по оцифровке печатных книг XVIII века, хранящихся в библиотеках Казани [6], потребовал решения комплекса проблем по устранению дефектов [5], сегментации [7], разработки специального драйвера клавиатуры [3], создания грамматических правил и словаря русского языка того времени.

Сложности многократно увеличиваются при попытке решения задачи распознавания текста рукописи. Имеется в виду так называемое оффлайновое распознавание, так как при онлайновом распознавании текста, вводимого в компьютер при помощи современных устройств рукописного ввода, в распоряжении программы имеется много дополнительной информации о процессе ввода, облегчающей задачу. Введение в электронное использование рукописных исторических документов, хранящихся в архивах и библиотеках России, имеет огромное научное и культурное значение, так как каждая рукопись, в отличие от большинства книг, уникальна. Следует отметить большой массив официально-деловых и частных документов XVII – XVIII вв., написанных скорописью с характерными выносными буквами и многочисленными диакритическими знаками, множество канцелярских бумаг, дневников, писем, черновиков литературных произведений XIX и XX вв. Одной из наиболее трудных проблем является расшифровка стенограмм. Так, в настоящее время остается не расшифрованной часть стенографических записей Ф.М. Достоевского, созданных его женой А.Г. Достоевской (Снеткиной). Сложность данной задачи определяется следующими моментами:

- нет людей, владеющих стенографической записью, которой пользовалась А.Г. Достоевская, известен только учебник, по которому она училась;
- стенографист может использовать свои нестандартные обозначения, так как обычно он расшифровывает записи сам;

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

- в стенографической записи применяется метод пропуска гласных букв, объединяются в один значок наиболее часто встречающиеся сочетания букв или наиболее распространенные слоги;
- некоторые значки стенографической записи очень похожи, но могут иметь разное значение, существенное значение имеет размер символа и т. д.

Распознавание рукописных исторических документов стало в последние годы одним из бурно развивающихся научных направлений. Активно разрабатываются методы устранения дефектов и улучшения качества цифровых изображений рукописей [13], а также сегментации строк [12]. Поскольку сегментация символов в рукописных текстах часто оказывается затруднительной, предлагаются специальные алгоритмы распознавания слитных слов [8] и целых строк, основанные на скрытых марковских моделях [15] и случайных полях [10]. Большое внимание уделяется распознаванию древнегреческих текстов [14] и арабских рукописей [11]. Анализ публикаций и практических задач в области распознавания исторических документов позволяет сделать следующие выводы:

- не существует алгоритмов распознавания, которые могли бы одинаково успешно применяться к рукописям разных языков и эпох; все алгоритмы ориентированы на определенные типы рукописного письма;
- даже самые признанные алгоритмы не доведены до готовых программных систем распознавания рукописей, которые можно было бы использовать в отечественных проектах по оцифровке исторических документов;
- практические задачи, требующие распознавания текстов оцифрованных рукописей, в настоящее время столь часты и разнообразны, что путь разработки для каждой такой задачи специализированного алгоритмического и программного обеспечения оказывается невозможным; в результате единственным универсальным способом на сегодняшний день остается ручной набор текста в редакторе MS Word; не стоит доказывать, что такой способ крайне неэффективен.

Поэтому весьма актуальной выглядит задача создания достаточно универсальной программной системы для автоматизированного распознавания рукописей и других текстов исторических документов, для которых автоматическое распознавание оказывается пока невозможным. Предлагаемая система автоматизированной дешифровки рукописных источников с возможностью интеллектуальной поддержки принятия решений при наборе позволит существенно ускорить процесс перевода рукописного текста в текстовый файл и повысит его точность.

2 Описание работы автоматизированной системы

Блок-схема работы автоматизированной системы представлена на рис. 1, а также на детализирующих рисунках 2 и 3.

В целом весь этап работы с системой можно разбить на две части:

- создание оригинальной графики символов;

На данном этапе на основе информации о возможной графике символов и оригинальных изображений уже распознанных текстов формируется виртуальная клавиатура, состоящая из набора векторных символов, встречающихся в исходном материале, сохраненных как векторные шрифты, например, в формате SVG. Также виртуальная клавиатура содержит в себе сопоставление между оригинальными символами (графемами) и буквосочетаниями (лексемами). Данная информация может быть записана как конфигурационный файл виртуальной клавиатуры, например, Keyboard.cfg. С помощью данного конфигурационного файла и набора векторных шрифтов можно предоставить пользователю возможность набирать текст оригинальными символами и при этом параллельно формировать перевод текста. Данная система использует векторные шрифты вместо растровых из-за их масштабируемости и возможности сравнения векторных символов с оригинальными растровыми. На рис. 2 представлена схема процесса формирования виртуальной клавиатуры. Данный процесс состоит из следующих последовательно выполняемых подпроцессов:

.1. *предобработка*; поскольку для составления полного списка оригинальных символов может потребоваться извлечение символов из оригинальных изображений текстов, последние необходимо подвергнуть предварительной обработке с целью устранения лишнего шума и улучшения их качества;

.2. *сканирование*; обработанные изображения, полученные как результат выполнения предыдущего процесса, подвергаются сканированию; данное сканирование заключается в поиске и выделении отдельных графем на изображении так, чтобы каждому символу текста соответствовала графема; в результате выполнения данного процесса получается список первичных растров;

.3. *коррекция растров*; оператор просматривает первичные растры с целью выявить какие-либо ошибки и устраняет их; в результате получается список растровых символов;

.4. *векторизация*; растровые символы анализируются и из них формируются векторные аналоги; на выходе – список первичных шрифтов;

.5. *коррекция шрифтов*; оператор просматривает список первичных шрифтов и исправляет обнаруженные дефекты; в результате получается список шрифтов;

.6. *сопоставление*; на основе набора шрифтов, информации о возможной графике символов и распознанных оригинальных текстов оператор создает таблицу соответствия между векторными символами и лексемами обычного текста;

.7. *формирование виртуальной клавиатуры*; на основе таблицы соответствия оператор с помощью данной системы формирует конфигурационный файл клавиатуры Keyboard.cfg.

- набор оригинального текста;

Набор текста осуществляется оператором с помощью созданной ранее виртуальной клавиатуры. В процессе набора символов система формирует и отображает пользователю список вариантов текущего набираемого слова, а также варианты перевода данного слова. Кроме того, система проверяет набранный текст на наличие синтаксических или стилистических ошибок и сообщает пользователю информацию об ошибках. В результате выполнения данного процесса система формирует текст, набранный оригинальными символами, а также перевод данного текста. На рис. 3 представлена схема процесса набора текста. Данный процесс состоит из следующих последовательно выполняемых подпроцессов:

.1. *предобработка*; оригинальное изображение перед непосредственным набором проходит предварительную обработку с целью устранения нежелательного шума и повышения его качества;

.2. *набор текста*; обработанное изображение подается в модуль набора текста системы, и на основании данного изображения оператор выполняет набор текста; в процессе набора система предлагает пользователю различные варианты набора и сообщает ему об ошибках; на основе этой информации пользователь решает, что делать дальше с текущим словом; после всех операций набора, выбора вариантов и корректировки на выходе формируется набранный оригинальными символами текст в виде изображения и текстовый перевод;

.3. *интеллектуальный набор*; система предлагает пользователю возможные варианты набора на основе введенных символов; выход данного процесса, варианты слова, подаются на вход основному процессу, набору текста; таким образом, обеспечивается обратная связь, используемая для коррекции набранного текста;

.4. *интеллектуальный перевод*; система анализирует графическое написание целого слова (а не отдельных букв) и предлагает пользователю варианты его дешифровки (перевода);

.5. *проверка правильности*; система проверяет синтаксис и стилистику текста и сообщает пользователю о выявленных ошибках.

3 Описание интерфейса модуля набора текста

Интерфейс программы можно разделить на 3 области (см. рис. 4):

1. Область рукописного текста, которая состоит из 2-х наложенных друг на друга слоев:
   a. нижний слой, содержащий оригинальное изображение;
   b. верхний слой, содержащий набранные символы.

Прозрачность обоих слоев подбирается так, чтобы набранные символы были отчетливо видны на фоне оригинальных символов. Фон верхнего слоя выбирается прозрачным, чтобы не накладываться на оригинальное изображение.

Кроме набранных символов верхний слой содержит также плавающий курсор (символьный курсор), указывающий место расположения текущего набираемого символа. Суть ввода заключается в ручном перемещении курсора через специальные клавиши либо попиксельно, либо перемещаясь по уже набранным символам также через специальные клавиши. Таким образом, перемещение курсора будет занимать только одну руку оператора, оставляя вторую свободной для набора символов.

2. Область печатного или распознанного текста, представляющая собой текстовое поле, содержащее соответствующие лексемы набранных символов, организованные в слова. По мере ввода символов программа автоматически вводит расшифрованные биграммы в текстовое поле на соответствующее место (как правило, в конец). Для текстового поля существует свой курсор (текстовый курсор), перемещение которого строго соответствует логическим перемещениям символьного курсора, т. е. текстовый курсор выполняет перемещение только в том случае, когда символьный курсор перемещается между символами. Пиксельные сдвиги символьного курсора не учитываются. Каждый последующий ввод символа за исключением пробела добавляет к текущему слову соответствующую биграмму или другой набор символов, соответствующий введенному символу. Область печатного текста логически выделяет текущее слово и по запросу пользователя осуществляет поиск вариантов слов из специального словаря, соответствующих текущему слову (как правило, совпадающих либо имеющих различие в одном символе; сравнение слов осуществляется посимвольно, с начала). Пользователь может посмотреть варианты слова в выпадающем списке, расположенном ниже текущего слова, с помощью специальной клавиши (см. рис. 2). Выбор слова из предложенных вариантов осуществляется стрелками.

При выборе нужного слова программа автоматически определяет недостающие символы и добавляет их в область рукописного текста, а также исправляет ошибочные символы (если выбранное слово отличалось от текущего одним символом). После этого в конец рукописного текста добавляется пробел, и курсор перемещается на новое место. Далее пользователь может, перемещая курсор, редактировать набранные символы по отдельности, изменяя их расположение и размер или непосредственно изменяя символы. Дальнейший ввод будет добавлять расшифрованные биграммы в новое текущее слово, и так далее, пока весь текст не будет расшифрован.

3. Панель наборных символов, состоящая из нескольких рядов кнопок с изображенными символами ввода. В случае большого количества наборных символов и недостаточного размера панели возможна прокрутка панели к нужному месту, чтобы сделать доступным недостающий набор символов. Кроме специальных символов ввода панель содержит пробел для логического разделения слов.

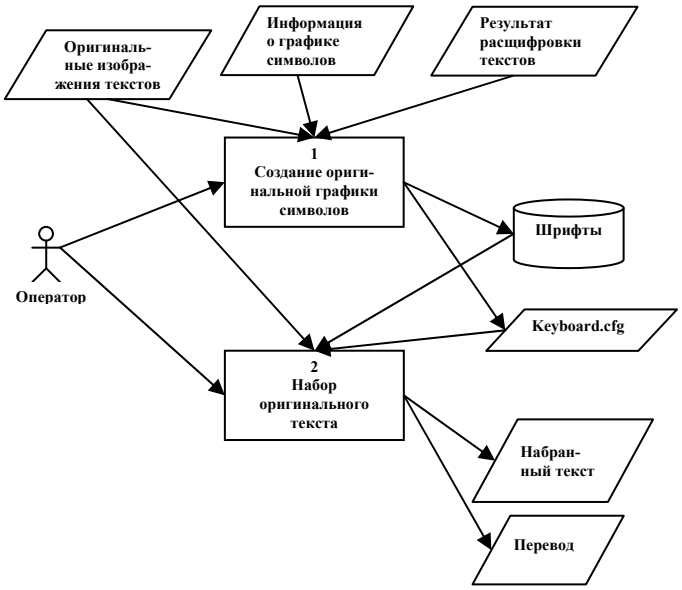

Рис. 1. Схема работы автоматизированной системы

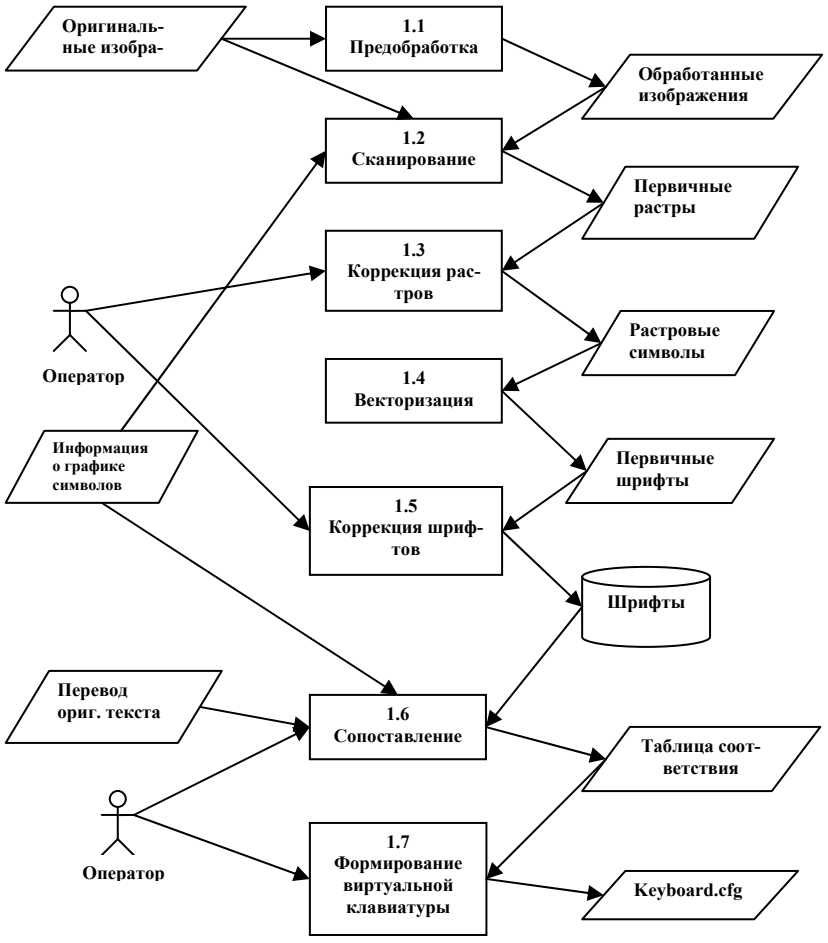

Рис. 2. Схема формирования виртуальной клавиатуры

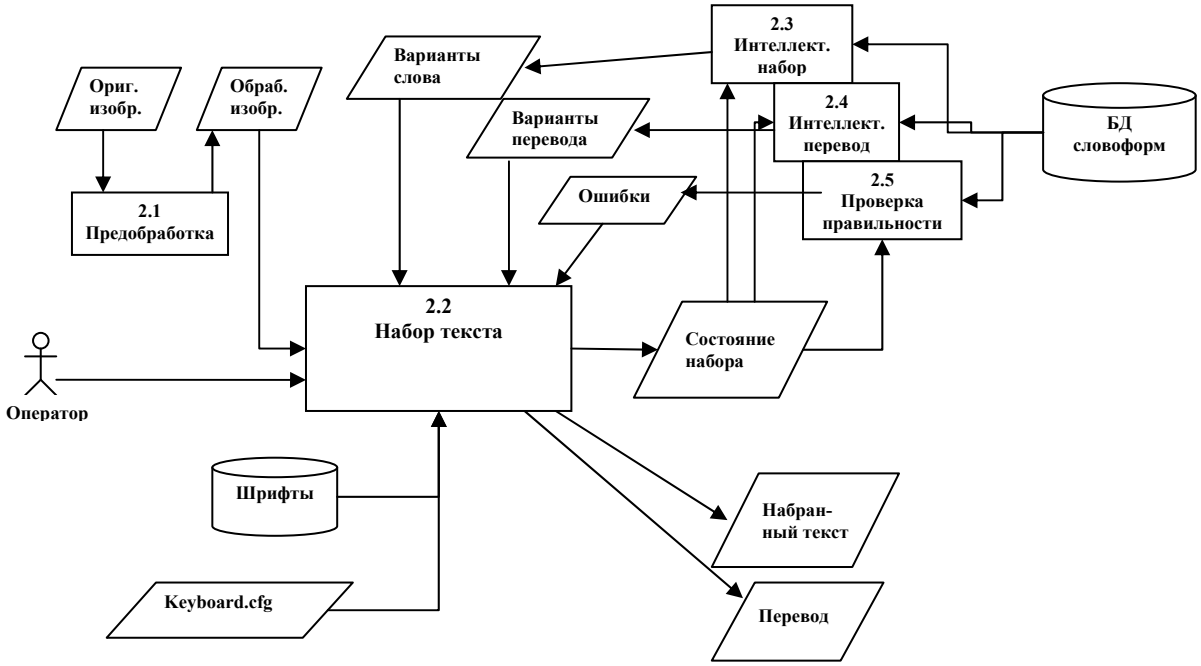

Рис. 3. Схема набора текста

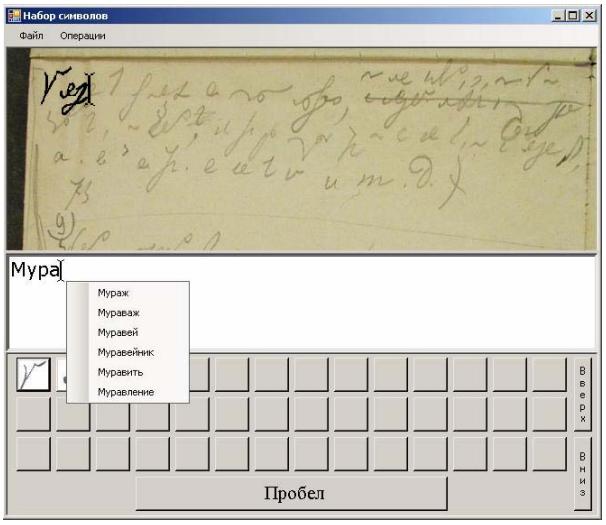

Рис. 4. Интерфейс модуля набора текста. Пример ввода символов с выпадающим списком вариантов слов

Ввод рукописного текста осуществляется путем комбинированного выбора символов (с помощью кликов мышкой) и позиционирования символьного курсора (с помощью специальных клавиш). Панель символов ввода формируется динамически с помощью заранее определенного набора символов (дефолтный набор). Дефолтный набор состоит из глифов символов ввода и соответствующих им символов печатного текста в виде биграмм или других буквосочетаний. Программа формирует матрицу кнопок с динамической привязкой к дефолтному набору символов, что позволяет использовать множество различных алфавитов для ввода. В процессе работы программы при каждом вводе рукописного символа программа с помощью привязки определяет пару <рукописный символ, буквосочетание>, соответствующую нажатой кнопке, и затем добавляет элементы этой пары в соответствующий текст.

4 Особенности и преимущества автоматизированной системы

Разрабатываемая программная система будет обладать следующими особенностями:

- ускоренный набор по сравнению с обычным набором;
- связь графического изображения текста и его текстового представления;
- интеллектуализация процесса набора, которая включает:
  a. поиск набираемого слова среди слов словаря и предложение возможных вариантов набора;
  b. анализ графического написания слова и предлагаемый вариант его дешифровки (анализ по графическому написанию целого слова, а не его отдельных букв).
- возможность организации виртуальной клавиатуры символов, что позволит набирать текст с помощью тех символов, которые использовались при создании источника;
- возможность автоматического распознавания в тексте отдельных часто повторяющихся фраз или слов путем поиска похожих фрагментов изображения, что существенно ускорит процесс обработки текста; кроме того, даже выявление нескольких ключевых слов в изображении документа позволит сделать его доступным для поисковых машин;
- возможность совместной работы по распознаванию больших коллекций текстов коллективом

исследователей с единым словарем и перекрёстной проверкой результатов.

Блок интеллектуального перевода набранного текста на современный язык будет содержать модули:

- распознавания рукописи, то есть перевода ее в текстовый формат в графике и орфографии оригинала;
- перевода языка оригинала на современный русский (или какой-либо другой) язык.

Модульная структура блока в случае необходимости позволит добавлять другие модули по переводу текста на современный язык. Работа данного блока должна основываться на словаре. В нашей системе по умолчанию используется база данных слов русского языка XIX века, содержащая более 50 000 словоформ. Она содержит написание слов в современной графике и графике XIX века, частоту употребления словоформ и т. д.

5 Некоторые особенности реализации системы

Предлагаемая нами автоматизированная система может быть реализована как настольное приложение. Но больший интерес, на наш взгляд, вызывает возможность ее реализации в виде веб-приложения. Это позволит, во-первых, добиться ее наибольшей универсальности и независимости от аппаратного и программного обеспечения (так как веб-браузеры есть на всех компьютерах, подключенных к интернету), и во-вторых, организовать совместную, распределенную работу сообщества исследователей над изображениями документов. Совместные исследования текстовых коллекций в рамках сетевых сообществ уже рассматривались ранее [1], однако использование подобных технологий на этапе расшифровки текстов выглядит не менее важным, так как позволяет одновременно повысить скорость распознавания и его точность за счет организации перекрёстной проверки.

Особое внимание хотелось бы уделить вопросу разработки шрифтов и их использования при наборе текста. Мы предлагаем использовать для этого передовую технологию SVG-шрифтов [16], которая, как ни странно, до сих пор очень редко используется для адекватного отображения документов в цифровых библиотеках [2]. Символы SVG-шрифтов описываются с помощью векторных графических примитивов в текстовом XML-формате. Существуют программы для разработки таких шрифтов и их конвертации из существующих TrueType-шрифтов. Современные версии браузеров Opera и Mozilla отображают SVG-шрифты без установки дополнительных плагинов. Наша автоматизированная система должна включать в себя редактор SVG-шрифтов и виртуальной клавиатуры, а также обеспечивать введение и отображение символов таких шрифтов поверх исходного изображения текста.

Литература

[1] Варфоломеев А.Г. Алгоритмическое и программное обеспечение работы сетевых сообществ исследователей текстов // Компьютерные науки и технологии. Ч. 1. Сборник трудов первой Межд. науч.-техн. конф., Белгород, 2009. – С. 30-33.

[2] Варфоломеев А.Г., Кравцов И.В., Филатов В.О. SVG-визуализация в цифровых библиотеках рукописных документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Девятой Всерос. науч. конф. RCDL'2007 (Переславль-Залесский, Россия, 14 – 18 октября 2007 г.). – Переславль-Залесский, 2007. – С. 230-235.

[3] Маргулис И.С. Раскладка клавиатуры для работы с русскоязычными текстами XVIII – XIX веков // Исследования по информатике. – 2005. – Вып. 9. – С. 133-138.

[4] Полное собрание законов Российской империи. Собрание первое, под ред. М.М. Сперанского, 1830 г. Формат DJVU. – http://www.pszri.ru.

[5] Соловьев В.Д., Южиков В.С. Автоматизированная система обработки и реставрации изображений старопечатных текстов и рукописей // Вестник КГТУ им. А.Н. Туполева. – 2006. – Вып. 3. – С. 28-30.

[6] Соловьев В.Д. Проект «Электронная библиотека книг XVIII века» – первый этап // Исследования по информатике. – 2005. – Вып. 9. – С. 117-124.

[7] Южиков В.С. Сегментация изображений страниц древних рукописей // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Девятой Всерос. науч. конф. RCDL'2007 (Переславль-Залесский, Россия, 14 – 18 октября 2007 г.). – Переславль-Залесский, 2007. – С. 236-240.

[8] Adamek T., O'Connor N.E., Smeaton A.F. Word matching using single closed contours for indexing handwritten historical documents//Int. J. on Document Analysis and Recognition. – 2007. – V. 9, No 2 – 4. – P. 153-165.

[9] Codices Electronici Sangallenses (CESG). – http://www.cesg.unifr.ch.

[10] Feng S., Manmatha R., Mccallum A. Exploring the use of conditional random field models and HMMs for historical handwritten document recognition// 2nd Int. Conf. on Document Image Analysis for Libraries (DIAL), 2006. – P. 8-37.

[11] Lorigo L.M., Govindaraju V. Offline Arabic handwriting recognition: a survey// IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2006. – V. 28, No 5. – P. 712-724.

[12] Malleron V. et al. Text lines and snippets extraction for 19th century handwriting documents layout analysis// Int. Conf. on Document Analysis and Recognition. – Barcelone, 2009. – P. 1001-1005.

[13] Moghaddam R.F., Cheriet M. Low quality document image modelling and enhancement// Int. J. on

Document Analysis and Recognition. – 2009. – V. 11, No 4. – P. 183-201.
[14] Ntzios K. et al. An old Greek handwritten OCR system based on an efficient segmentation-free approach//Int. J. on Document Analysis and Recognition. – 2007. – V. 9, No 2 – 4. – P. 179-192.
[15] Plötz T., Fink G.A. Markov models for offline handwriting recognition: a survey//Int. J. on Document Analysis and Recognition. – 2009. – V. 12, No 4. – P. 269-298.
[16] Scalable Vector Graphics (SVG) 1.1 Specification. W3C Recommendation 14 January 2003. Part 20. Fonts. – http://www.w3.org/TR/SVG11/fonts.html.

Automated recognition system for handwritten historical documents

A.A. Rogov, A.N. Talbonen, A.G. Varfolomeev

This article is devoted to such issue as recognition of handwritten historical documents such as F.M. Dostoyevsky's records. The article contains description of intended automated decyphering system for handwritten sources, as well as the basic processes associated with handwritten text typing.