На правах рукописи

Зеленцов Иван Анатольевич

МЕТОДИКА РАСПОЗНАВАНИЯ ДРЕВНЕРУССКИХ СКОРОПИСНЫХ ТЕКСТОВ

05.13.17 — Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата технических наук

Москва – 2011

Работа выполнена на кафедре Систем обработки информации и управления Московского государственного технического университета им. Н.Э. Баумана

Научный руководитель: Кандидат технических наук, доцент **Филиппович Юрий Николаевич**

Официальные оппоненты: Доктор технических наук, профессор **Моттль Вадим Вячеславович**
Кандидат физико-математических наук, доцент **Варфоломеев Алексей Геннадьевич**

Ведущая организация: Институт проблем информатики РАН

Защита диссертации состоится 09 февраля 2012г. в 16 часов 30 минут на заседании диссертационного совета Д 212.141.10 при Московском государственном техническом университете им. Н.Э. Баумана.

С диссертацией можно ознакомиться в библиотеке Московского государственного технического университета им. Н.Э. Баумана.

Автореферат разослан «__»_____2011г.

Учёный секретарь
диссертационного совета
кандидат технических наук, доцент Иванов С.Р.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность исследования. В настоящее время в архивах и библиотеках накоплено большое количество древнерусских рукописей различных временных периодов. Одним из классов таких документов являются скорописные тексты XVII в. Для обеспечения возможности компьютерного анализа, хранения и электронного переиздания этих документов требуется их перевод в электронный вид. Значительный объем задачи, а также весьма узкий круг специалистов, в числе которых ученые палеографы, историки, лексикографы и филологи, обладающих знаниями и навыками чтения скорописных рукописей, порождают необходимость в автоматизации данных процессов.

Сложность автоматизации получения электронных текстовых представлений скорописных рукописей обусловлена, прежде всего, спецификой используемого в них языка и стиля письма, а также их ветхостью. Эти факторы затрудняют использование применительно к рассматриваемым документам существующих средств распознавания текста, ориентированных на современные языки и способы представления текстовой информации на бумажных носителях.

Таким образом, актуальной является задача разработки методики автоматизированного распознавания, учитывающей особенности скорописного способа формирования текста, применявшегося в России XVII в.

Целью исследования является разработка методики распознавания, позволяющей осуществлять автоматизированный перевод древнерусских скорописных текстов XVII в. из растровых изображений в вид электронных текстов.

Задачи исследования:

1. изучение особенностей древнерусской скорописи XVII в.;
2. анализ существующих методов распознавания;
3. разработка подхода к решению задачи распознавания скорописи;
4. разработка способа структурного описания объектов распознавания и метода их формального представления;
5. разработка алгоритмов распознавания;
6. теоретическое и практическое исследование предложенных алгоритмов.

Объектом исследования является древнерусский скорописный текст XVII в. с точки зрения его компьютерного распознавания.

Предметом исследования выступает методика распознавания скорописных текстов XVII в.

Научную новизну диссертационного исследования составляют следующие полученные результаты:

1. Предложена методика распознавания древнерусской скорописи XVII в.
2. Предложен подход к распознаванию древнерусских скорописных текстов XVII в., основанный на реконструкции начертаний символов текста

с использованием экспертных палеографических знаний на этапе обучения и управляемый двухуровневой интерактивной архитектурой «буква-слово» проверки гипотез на этапе распознавания.

3. Предложен способ описания структур букв и слов, а также метод представления их структурных описаний на основе фреймовых сетей, отличающийся рекурсивностью описаний объектов различных структурных уровней и совместным использованием структурных элементов в описаниях схожих объектов.

4. Разработаны алгоритмы распознавания букв и слов скорописи путём выдвижения и проверки гипотез относительно распознаваемых объектов. Отличительными особенностями алгоритмов являются применение динамических фреймовых структур для описания распознанных фрагментов изображения и представление гипотез в виде схем согласования динамических фреймов с фреймами базы знаний.

Методы исследования. В работе использованы методы теории множеств, дескриптивной логики, комбинаторики, нечёткой логики. При решении практических задач использован объектно-ориентированный подход к построению программных систем.

На защиту выносятся научные положения, составляющие научную новизну исследования.

Достоверность и обоснованность научных положений обеспечивается корректностью применения математического аппарата при построении и исследовании моделей и алгоритмов, а также подтверждается результатами экспериментальных исследований предложенных алгоритмов. Предложенная методика апробирована на конференциях и в научных публикациях.

Практическая значимость. Диссертационное исследование направлено на развитие технических средств, используемых в культурно значимых исследованиях памятников письменности. Предложенная в работе методика может быть использована при построении систем автоматизированного перевода имеющихся фондов скорописных документов в электронное текстовое представление. Подобные программные средства предназначаются для научных сотрудников, проводящих лингвистические исследования древних документов данного вида, а также для специалистов, участвующих в создании электронных хранилищ памятников письменности и подготовке их мультимедийных электронных изданий.

Практическая ценность. Использование компьютерных программных средств, построенных на основе предложенной методики распознавания, позволит сократить затраты времени на получение электронных текстовых версий документов за счёт замены этапа ручного ввода автоматизированным распознаванием. Наличие электронных текстовых версий скорописных документов делает доступным применение к ним всевозможных компьютерных технологий по обработке и анализу текста.

Использование результатов работы. Материалы проведенного исследования были использованы: в учебном процессе кафедры Систем обработки информации и управления МГТУ им. Н.Э.Баумана, при чтении лекций и курсовом проектировании по дисциплине «Лингвистическое обеспечение АСОИУ»; в учебном процессе кафедры Медиасистем и технологий МГУП им. Ивана Федорова в заданиях производственной практики; в научном исследовании по гранту Президента РФ для государственной поддержки молодых российских ученых – кандидатов наук МК-3732.2010.9 «Разработка словарных компонентов интегрированной информационной технологии переиздания печатных источников XVIII – нач. XIX вв.»; в научных исследованиях древнерусской языковой культуры ученых и специалистов ИРЯ им В.В.Виноградова РАН, Российской государственной библиотеки.

Результаты диссертационного исследования размещены в сети Интернет по адресу http://it-claim.ru/Projects/Skoropis/SkoropisBooks.htm, в их числе представлены 4 древнерусские скорописные книги XVII в., снабженные графическими справочниками составляющих эти документы графем и начертаний словоупотреблений.

Апробация работы. Основные результаты диссертационной работы докладывались на заседаниях комиссии по аттестации аспирантов кафедры Систем обработки информации и управления МГТУ им. Н.Э. Баумана в 2009-2011 гг. Материалы работы были также представлены на следующих научных конференциях и семинарах: Научной школе для молодых учёных «Компьютерная графика и математическое моделирование (Visual Computing)» (г. Москва, 2009); Научной межвузовской конференции преподавателей, аспирантов, молодых учёных и специалистов «Печатные средства информации в современном обществе» (г. Москва, 2010); Международной научной конференции «Информационные технологии и письменное наследие El'Manuscript-10» (г. Уфа, 2010); Научно-методических семинарах и вебинарах НОК CLAIM (г. Москва, 2008-2010, URL: http://it-claim.ru/Education/Seminar/SeminarCLAIM.htm); Научно-технической международной молодежной конференции «Системы, методы, техника и технологии обработки медиаконтента» (г.Москва, 2011).

Публикации по теме диссертации. Основные результаты по теме диссертации опубликованы в 7-и печатных работах, в том числе 3-х – в журналах, включённых в перечень ВАК РФ. Электронные версии всех печатных публикаций представлены в Интернет по адресу: http://it-claim.ru/Projects/Skoropis/Skoropismain.htm.

Объём и структура работы. Диссертация состоит и списка терминов и сокращений, введения, четырёх глав, заключения, списка использованных источников из 97 наименований и 5 приложений. Основной текст изложен на 174 страницах, включающих 45 рисунков и 14 таблиц. Приложения выполнены на 33 страницах.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении к диссертации обосновывается актуальность работы, определена цель исследования и перечислены его задачи, сформулирована научная новизна исследования, практическая значимость и ценность его результатов.

В первой главе диссертации приводится описание объекта и предмета исследования, определяются подлежащие решению задачи и критерии качества системы распознавания древнерусской скорописи XVII в., приводится обзор существующих методов и систем распознавания, предлагается подход к решению задачи распознавания скорописи.

Рассматриваемые в работе вопросы, связанные с распознаванием рукописного текста, освещались ранее в работах Ю.И.Журавлёва, А.Б.Меркова, К.Фу, Д.Доэрманна (D.Doermann), М.Идена (M.Eden) и др.

Скоропись — форма кириллического письма, возникшая во второй половине XIV в. В XVI-XVII вв. скоропись господствует в области делового письма. В то же время область её применения расширяется: её начинают применять при переписке памятников литературного характера. Графически скоропись этого периода строится в значительной мере на основе полукруглых очертаний букв, в отличие от предшествующих форм письма: устава и полуустава (рис. 1).

В ходе анализа данного вида текстов выявлен ряд особенностей, затрудняющих его распознавание существующими системами OCR. Начертание букв сильно варьируется, буквы часто имеют соединительные элементы. В буквах могут отсутствовать предполагаемые пересечения линий и присутствовать дополнительные декоративные росчерки. Линии соседних букв могут пересекаться друг с другом в случайных местах и иметь дефекты.

Проведён анализ существующих методов распознавания текста и возможности их применения к древнерусской скорописи. Их классификация показана на рисунке 2.

Невозможность выделения отдельных букв и слов скорописи из изображения не позволяет использовать для описания объектов распознавания методы на основе Евклидова пространства и списков признаков, реализующие параллельную процедуру распознавания, т.е. выполняющие анализ всего образа единовременно. В связи с этим целесообразным является применение последовательного подхода к анализу изображения, реализуемого структурными методами распознавания.

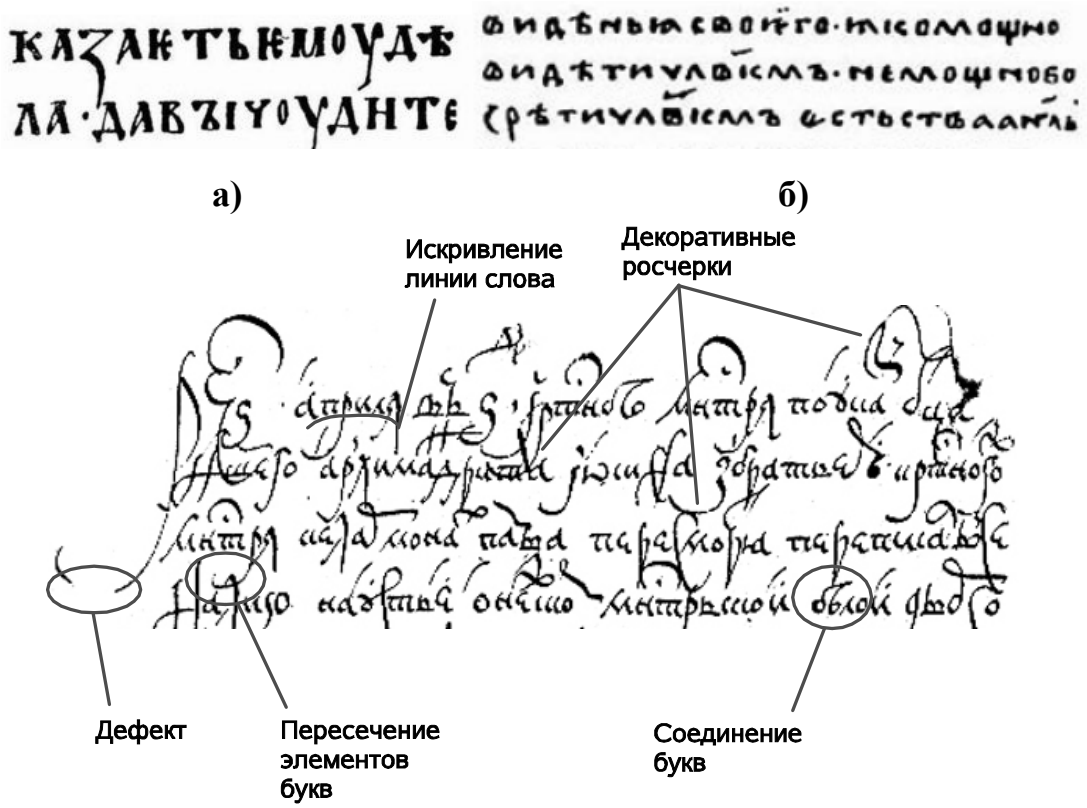


Рис.1. Иллюстрация особенностей скорописного формирования текста (а – устав; б – полуустав; в – скоропись)

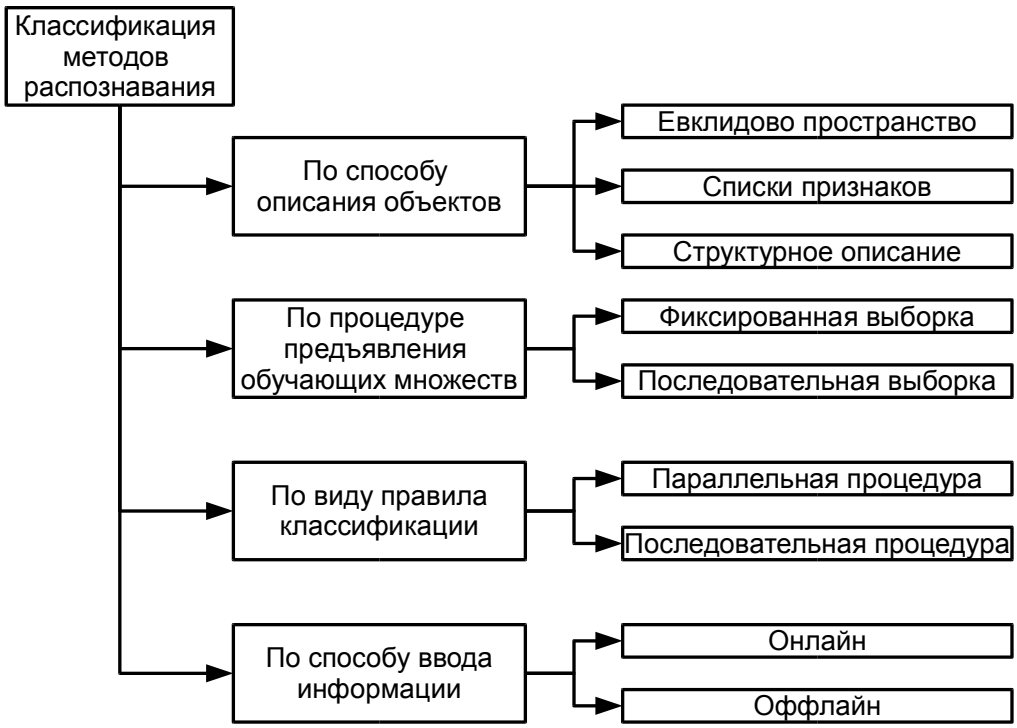


Рис.2. Классификация методов распознавания

Проведён анализ существующих систем распознавания с точки зрения решения поставленной задачи и особенностей скорописи XVII в. (табл. 1). На его основе делается вывод, что большинство допущений относительно распознаваемого текста, принимаемых современными системами OCR, не применимы к скорописным документам. Сделан вывод о том, что разработка специализированной методики распознавания является оправданной.

Таблица 1. Сравнение существующих средств распознавания текста

Названия систем распознавания	**Вариатив. начертаний букв**	**Пересеч. элементов букв**	**Декор. росчерки**	**Необх-ть выделения букв**	**Древнерус. скоропись XVIIв.**
FineReader, Form Xtra, Cognitive Forms	Отн. шрифта	Нет	Нет	Да	Нет
PenReader, «Рукопись», CalliGrapher	Да	Да	Нет	Нет	Нет
Необходимые возможности	Да	Да	Да	Нет	Да

В работе предлагается новый структурный подход к распознаванию, ориентированный на скорописные тексты XVII в., характеризуемый следующими особенностями (принципами):

1. *Трассировка (векторизация) линий, составляющих изображения букв.* Предполагается, что начертание символов скорописи производилось по определённым правилам. Поэтому в качестве структурных элементов рассматриваются траектории движения пера по бумаге, имевшие место при формировании текста (*реконструкция* процесса начертания символов).
2. *Нечёткое сравнение* описаний структурных элементов для решения проблемы варьируемости начертания символов.
3. *Распознавание под управлением* гипотез для преодоления связанности изображений букв и исключения из рассмотрения неинформативных декоративных элементов букв. Поскольку набор слов, использованных в скорописных документах, является априори известным, предлагается использовать словник для повышения качества распознавания. Распознавание текста представляется в виде распознавания всех его слов, а распознавание каждого слова заключается в распознавании составляющих его букв. Гипотезы в таком случае выдвигаются относительно слов текста и относительно букв, составляющих слова.
4. *Использование экспертных палеографических знаний* при формировании базы знаний о способах начертания символов.
5. *Интерактивность процесса распознавания.* Допущение наличия на изображении сложных участков, на которых для продолжения распознавания требуется вмешательство оператора.

Во второй главе описывается предлагаемая методика распознавания древнерусской скорописи, основанная на сформулированных в первой главе принципах. Предлагается способ структурного описания объектов распознавания и метод их представления в базе знаний системы.

На рис. 3 представлена структурная схема системы распознавания, реализующей предложенную методику. Компонентами системы являются: трассировщик; распознаватель, состоящий из распознавателей слов и букв; база знаний, содержащая информацию о буквах и словах; модуль обучения. Компонент *трассировщик* выполняет функцию выделения во входном изображении структурных элементов букв. Компонент *распознаватель* отвечает за структурное распознавание входного изображения и формирование выходного электронного документа. Его составляющие части, *распознаватель букв* (РБ) и *распознаватель слов* (РС), реализуют предложенную двухуровневую схему распознавания. *База знаний* (БЗ) содержит полученную от эксперта информацию о структуре распознаваемых букв, а также словник с описанием структуры каждого из слов. За получение этой информации от эксперта и наполнение базы знаний отвечает *модуль обучения*.

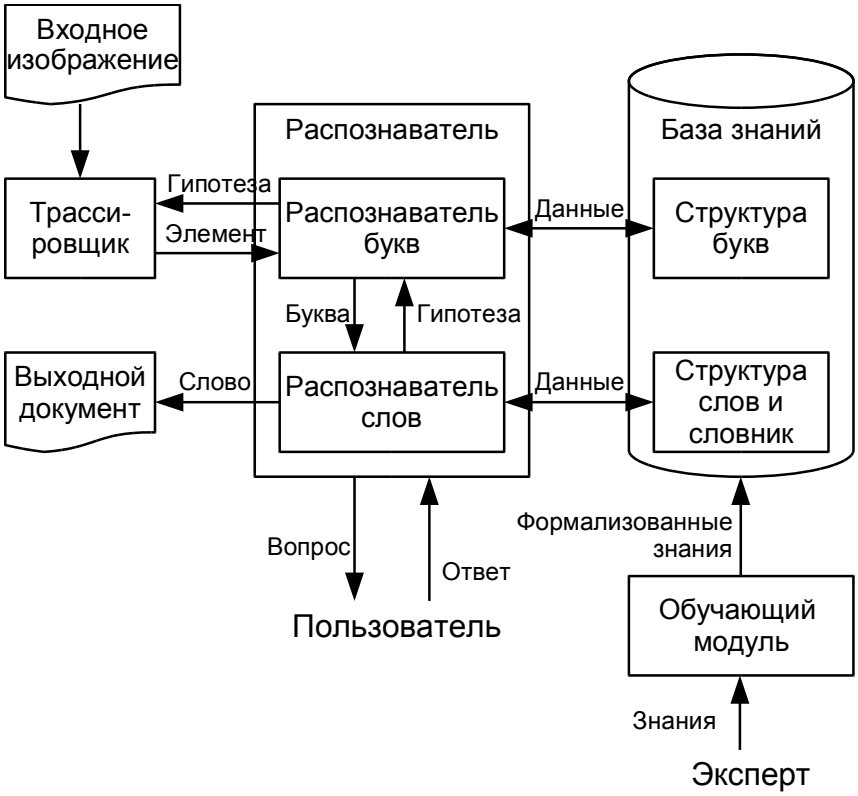


Рис.3. Структурная схема системы распознавания

Основной цикл распознавания заключается в серии распознаваний модулем РС всех слов входного документа поочерёдно. Для распознавания очередного слова РС выдаёт для РБ серию команд на распознавание очередной буквы. Запрос к РБ содержит информацию о прямоугольной области изображения, в которой должен осуществляться поиск, и код буквы – оп-

циональная предварительная гипотеза. На первом шаге распознавания слова предварительная гипотеза РС отсутствует, и команда заключается в запросе на распознавание любой буквы.

РБ, в свою очередь, управляет работой модуля трассировки изображения, запрашивая выделение на изображении очередного элемента буквы. Трассировщик выполняет поиск на основе информации о виде и местоположении ожидаемого элемента, если они указаны в запросе. В противном случае выполняется поиск элемента любого известного вида. Полученная информация о выделенном элементе возвращается в модуль РБ. Она содержит его точные геометрические параметры и местоположение. РБ анализирует информацию о геометрии элемента и заносит эти данные в строимую им модель изображения. Запрашивая в базе знаний набор букв, содержащих элемент найденного вида, РБ может выдвигать гипотезы о распознаваемой букве. Это позволяет ему действовать в режиме проверки выдвинутых гипотез, выдавая трассировщику запросы на поиск элементов ожидаемого вида с указанием их примерного местоположения относительно уже найденных элементов. Во время проведения серии проверок РБ может отклонять гипотезы и выдвигать новые. В результате, когда одна из гипотез подтверждается в достаточной степени, РБ приостанавливает свою работу и передаёт соответствующий код буквы в РС.

Получив от РБ код и местоположение распознанной буквы, РС обращается к словарной части базы знаний и получает набор слов, содержащих эту букву. Теперь можно принять одно из этих слов в виде гипотезы и в соответствии с ней предположить, какая буква будет найдена следующей. Эта буква помещается в следующий запрос к РБ и является для него начальной гипотезой. Результат выполнения этого запроса добавляется к полученным ранее результатам, и из списка гипотез удаляются слова, не содержащие соответствующую последовательность букв. Когда одна из гипотез признаётся в достаточной мере согласованной с наблюдаемой картиной, это слово передаётся управляющей части распознавателя, которая помещает его в выходной документ в кодированном виде.

Работа системы распознавания строится на интерактивном принципе. В случаях, когда система оказывается не в состоянии выполнить распознавание какого-либо фрагмента изображения в связи с его сложностью или наличием в нём крупного дефекта, она может выдать запрос на разрешение проблемы оператору, контролирующему процесс распознавания. При этом указываются граница проблемной области и выдвинутые к моменту останова гипотезы относительно наблюдаемого слова. Оператор может самостоятельно разобрать данный фрагмент текста, после чего указать системе точку, с которой она может продолжить распознавание.

Разработан способ структурного описания изображений букв. В качестве структурных элементов букв используются линии и точки их пересечения. Для описания линий различных видов применяются характеристики

пути линии (в виде цепного кода угловых измерений направлений обхода линии за 10 шагов) (рис. 4а) и *формы линии* (угол наклона диагонали описывающего прямоугольника линии, рис. 4б). Точка пересечения линии другой линией характеризуется её относительным положением по горизонтали и вертикали внутри описывающего прямоугольника линии (рис. 4в). Сравнение данных характеристик выполняется нечётким способом путём вычисления степени близости их числовых параметров.

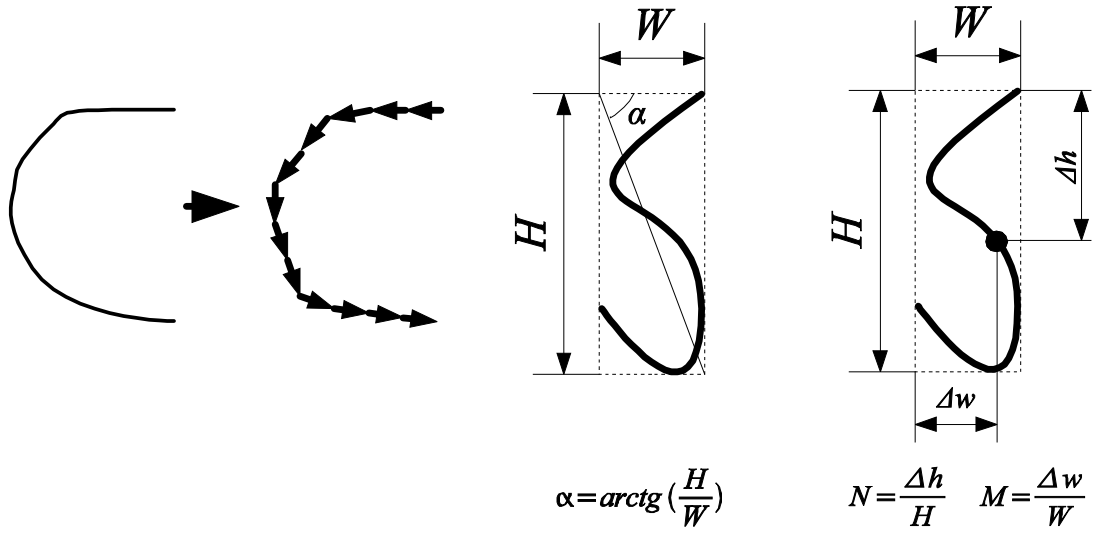


Рис.4. Описание структурных элементов букв
(а – путь линии; б – форма линии; в – положение точки пересечения)

Предложен метод представления структурных описаний объектов распознавания в базе знаний (БЗ) системы. В качестве метода представления знаний о структурах объектов распознавания использованы фреймовые сети. Сетевая природа фреймового представления позволяет корректно описывать сложный набор взаимосвязи структурных элементов букв. Кроме того, фреймовые сети по своей идее связаны с принципом проверки гипотез.

Разработана схема построения баз знаний. БЗ состоит из множества узлов, описываемых как $Б = (D, E, F) = (D, E, Q \cup R)$, где: $D = D_W \cup D_L$ – множество *ДетализируемыхУзлов*: *НачертанияСлов, НачертанияБукв*; $E = E_W \cup E_L \cup E_{LN} \cup E_{PT}$ – множество узлов-*Элементов*: *Слова, Буквы, Линии, Точки* (пересечения линий); F – множество *ВхожденийСвойств*: узлов, описывающих вхождения *Элементов* в структуру *ДетализируемыхУзлов*, а также их взаимоотношения; $Q \subseteq F$ – множество *ВхожденийЭлементов*: *ВхожденияБукв, ВхожденияЛиний, ВхожденияТочек*; $R \subseteq F$ – множество *ВхожденийОтношений*: *ПространственныеОтношения*, *ПринадлежностиТочек, СоответствияТочек*.

Основными типами узлов базы знаний являются *Слово* и *Буква* (рис. 5). Каждая буква может иметь несколько начертаний, что описывается связью каждого узла типа *Буква* с набором узлов типа *НачертаниеБуквы* (подтип *ДетализируемогоУзла*). Аналогично, *Слово* может иметь несколько *НачертанийСлова* (как правило, одно). С каждым *ДетализируемымУзлом*

связан набор узлов типа *ВхождениеСвойства*, описывающих структуру соответствующего начертания. Одним из видов *ВхожденийСвойств* являются узлы типа *ВхождениеЭлемента*; их назначением является указание (индикация) присутствия в структуре *ДетализируемогоУзла* конкретных *Элементов*, описывая структурный состав объектов. Узлы типа *ВхождениеОтношения* связывают *ВхожденияЭлементов* различными способами, определяя структурные отношения элементов объектов.

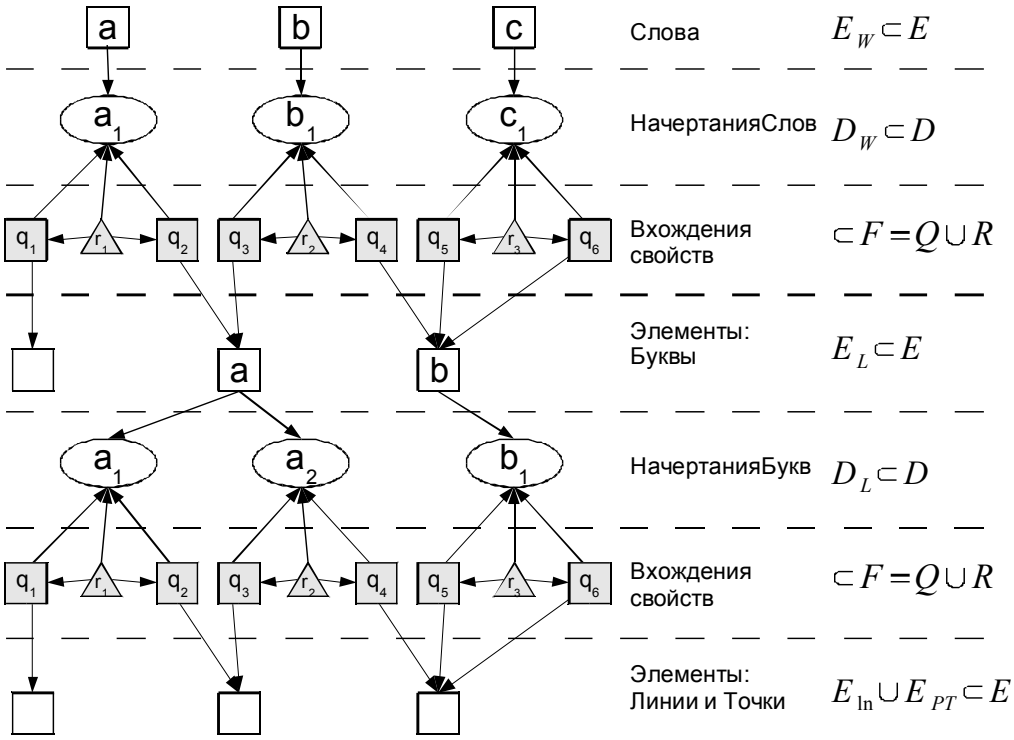


Рис.5. Схема построения базы знаний.

Элементами в структуре *НачертанийБукв* являются *Линии* и *Точки* (пересечения линий). *ВхожденияТочек* привязываются к содержащим их *ВхождениямЛиний* узлами типа *ПринадлежностьТочки*, соответствие *Вхождений Точек* друг другу задаётся узлами типа *СоответствиеТочек*. Взаимное расположение *ВхожденийЛиний* указывается узлами типа *ПространственноеОтношение. Элементами* в структуре *НачертанийСлов* являются *Буквы*. *ВхожденияБукв* связываются между собой *ПространственнымиОтношениями*.

На основе предложенной схемы построения БЗ введены количественные показатели содержимого БЗ в части описания букв и слов, которые позволили дать теоретическую оценку среднего числа узлов, необходимых для описания буквы и слова.

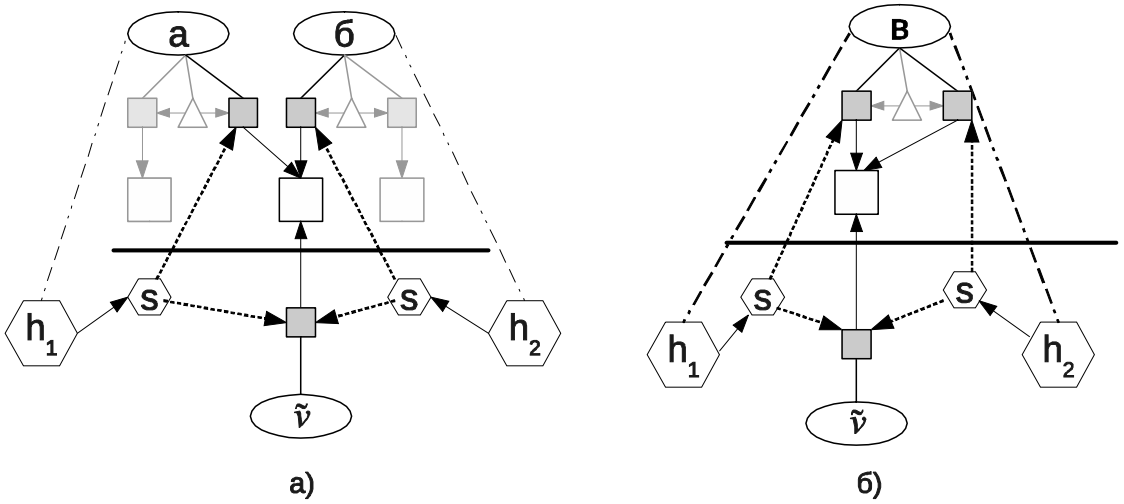
Сформулированы методики наполнения БЗ структурными описаниями указанного вида. Формирование в БЗ структурных описаний букв выполняет модуль обучения путём ведения диалога с обучающим систему экспертом-палеографом. Эксперту предлагается изобразить в специальной области ин-

терфейса модуля обучения скорописные начертания всех букв, которые должна распознавать система. Для этого он может пользоваться мышью, графическим планшетом или подобным устройством ввода. Модуль обучения производит онлайн-анализ вводимых изображений и формирует соответствующие фреймовые описания. Наполнение БЗ в части описания слов производится путём автоматического анализа заранее подготовленного файла, содержащего список распознаваемых слов в текстовом виде.

Третья глава посвящена описанию алгоритмов функционирования компонентов системы. Приводится описание использованного способа трассировки изображений. Формулируется алгоритм распознавания абстрактных образов, на основе которого конструируются алгоритмы распознавания букв и слов скорописи. Проводится теоретическое исследование эффективности данных алгоритмов

Назначением компонента «трассировщик» системы распознавания является выполнение реконструкции начертаний символов скорописи с целью выявления их структурных элементов. К настоящему моменту предложен ряд методов, позволяющих восстанавливать траекторию движения пишущего инструмента в оффлайн-изображении. Предлагаемая в диссертационном исследовании методика распознавания формулирует ряд требований к порядку функционирования модуля трассировки, оставляя за рамками выбор и обоснование конкретной его реализации, а также методики предобработки входных изображений. В качестве рабочего варианта в исследовании использован способ трассировки, основанный на истончении линий изображения.

Разработан алгоритм распознавания букв и слов на основе концепции Виртуального фрейма. В процессе распознавания образа для сохранения получаемой информации об изображении в динамической памяти системы строится фреймовая модель, описывающая наблюдаемую в каждый текущий момент картину – *виртуальный фрейм* (ВФ). Он строится по тем же правилам, что и фреймы распознаваемых объектов в БЗ и служит для сопоставления наблюдаемой картиной с фреймами БЗ. Задача распознавания сводится к нахождению способа установления соответствия между узлами виртуального фрейма и узлами одного из фреймов букв в базе знаний. В каждый момент процесса распознавания состав и структура виртуального фрейма позволяет выделить набор фреймов базы знаний в качестве списка гипотез, потенциально описывающих наблюдаемую картину. С нахождением новых элементов на изображении этот список будет сокращаться. Для описания выдвинутых гипотез для каждой из них строится специальная структура, состоящая из пар ссылок на согласованные узлы и тем самым описывающая схему согласования. Под *гипотезой* понимается пара *(h, S(h)),* где: *h* – динамический узел, указывающий предполагаемый гипотезой в качестве ответа распознавания фрейм в БЗ; *S(h)* – набор пар ссылок, указывающий на согласуемые данной гипотезой пары узлов ВФ и предполагаемого фрейма БЗ (рис. 6).



а) б)

Рис. 6. Виртуальный фрейм и гипотезы (на рисунке: $\tilde{v}$ - виртуальный фрейм; h_1,h_2 – узлы гипотез; s – узлы пар согласований из $S(h)$).

Предложен ряд формальных характеристик гипотез, конструируемых указанным способом, позволяющих признавать гипотезы подтверждёнными или отвергнутыми. Ими являются: степень согласованности, степень пригодности, степень проверенности и степень успешности, измеряемые в интервале [0;1]:

$$N^{согл}(h)=\frac{w(h)}{w\left(F(d)\right)};\qquad N^{приг}(h)=\begin{cases}\dfrac{w(h)}{w\left(F(\tilde{v})\right)}, & w\left(F(\tilde{v})\right)>0\\ 1, & w\left(F(\tilde{v})\right)=0\end{cases};$$

$$N^{пров}(h)=\frac{w_{П}\left(Q^{пров}(h)\right)}{w_{П}\left(Q(d)\right)};\qquad N^{усп}(h)=\frac{w_{П}\left(S_{Q}^{БЗ}(h)\right)}{w_{П}\left(Q^{пров}(h)\right)}.$$

Здесь: $w(h), w\left(F(d)\right), w\left(F(\tilde{v})\right)$ – соответственно подсчёты количеств согласований в гипотезе h и узлов-*Вхождений* в фрейме БЗ d и виртуальном фрейме $\tilde{v}$, взвешенные по типам узлов; $w_{П}\left(Q^{пров}(h)\right), w_{П}\left(Q(d)\right), w_{П}\left(S_{Q}^{БЗ}(h)\right)$ – подсчёты количеств узлов-*ВхожденийЭлементов*, проверенных в рамках гипотезы h, узлов- *ВхожденийЭлементов* в фрейме БЗ d, согласований *ВхожденийЭлементов* в гипотезе h, взвешенные по типам проверяемых узлов.

Сформулирован и доказан ряд утверждений, описывающих законы изменения значений данных характеристик в процессе распознавания.

На основе предложенного механизма построения ВФ и построения гипотез, а также указанных утверждений, сформулирован алгоритм распознавания абстрактных образов, работа которого заключается в следующем. На изображении отыскивается элемент одного из известных видов. На его основании выдвигается список первоначальных гипотез. Далее на каждой итерации цикла одна из гипотез принимается текущей. На основе её текущего состояния предсказывается положение и вид очередного структурного элемен-

та и выполняется проверка этого предположения. Результат проверки заносится в виртуальный фрейм, после чего все имеющиеся гипотезы согласуются с поступившей информацией. Далее, на основе вычисления описанных выше характеристик, из списка гипотез удаляются опровергнутые, а также, возможно, выдвигаются новые. Работа прекращается, когда не остаётся непроверенных гипотез.

Теоретически показано, что данный алгоритм сходится к правильному ответу в случае успешного завершения; оценено число шагов алгоритма в зависимости от количественных характеристик содержимого БЗ.

На основе алгоритма распознавания абстрактных образов сформулированы конкретные алгоритмы распознавания букв и слов, теоретически доказана корректность данных алгоритмов и установлены зависимости времени выполнения алгоритмов от содержимого используемых баз знаний. Число шагов алгоритма распознавания букв и слов базы знаний максимально оценивается как $\widetilde{c}_{\max} = \frac{\widetilde{n}_{БЛ}^2 |D_L|}{|E_{LN}|}$, $\widetilde{c}_{\max} = \frac{\widetilde{n}_{СБ}^2 |D_W|}{|E_L|}$., где: $\widetilde{n}_{БЛ}$, $\widetilde{n}_{СБ}$ – среднее количество линий в начертании буквы и букв в слове соответственно; $|D_L|$, $|D_W|$ – количество описаний начертаний букв и букв в слове в базе знаний; $|E_{LN}|$, $|E_L|$ – количество линий различных видов, использованных в БЗ для описания начертаний всех букв и количество букв алфавита.

Четвертая глава посвящена вопросам, связанным с программной реализацией предложенной методики и алгоритмов распознавания. Приводится описание реализации компонентов системы распознавания. Рассматриваются эксперименты, проведённые над реализованными модулями с целью исследования эффективности методики и проверки теоретических расчётов.

Для проведения описанных ниже исследований был реализован экспериментальный программный комплекс, содержащий модули распознавания букв и слов, модуль обучения, модуль управления базой знаний (реализованы на языке Java), а также модуль трассировки (язык C++, библиотека манипулирования изображениями IPL). В качестве технологии описания содержимого БЗ использован язык OWL; программный доступ к OWL-базам осуществлён посредством библиотеки Jena.

Целью исследования эффективности алгоритмов распознавания слов и букв является проверка полученных в третьей главе теоретических оценок времени выполнения алгоритмов. Были подготовлены наборы баз знаний, в которых варьировались количественные характеристики, входящие в теоретические выражения критерия эффективности (числа шагов алгоритмов). При исследовании алгоритма распознавания букв проводились серии распознаваний массивов изображений отдельных букв с использованием указанных БЗ. При проверке алгоритма распознавания слов проводились серии распознавания слов с эмуляцией «идеального» распознавания букв.

Полученные результаты подсчёта числа шагов алгоритмов показывают, что установленный характер зависимостей эффективности алгоритмов от количественных показателей БЗ подтверждается на практике. Кроме того, установлено, что теоретические максимальные оценки числа шагов для распознавания среднего по БЗ образа не превышаются в большинстве случаев.

Проведено исследование алгоритма распознавания букв с точки зрения его корректности путём распознавания наборов изображений отдельных букв. В качестве критерия использовался процент успешных распознаваний. Результаты исследования представлены на рисунке 7.

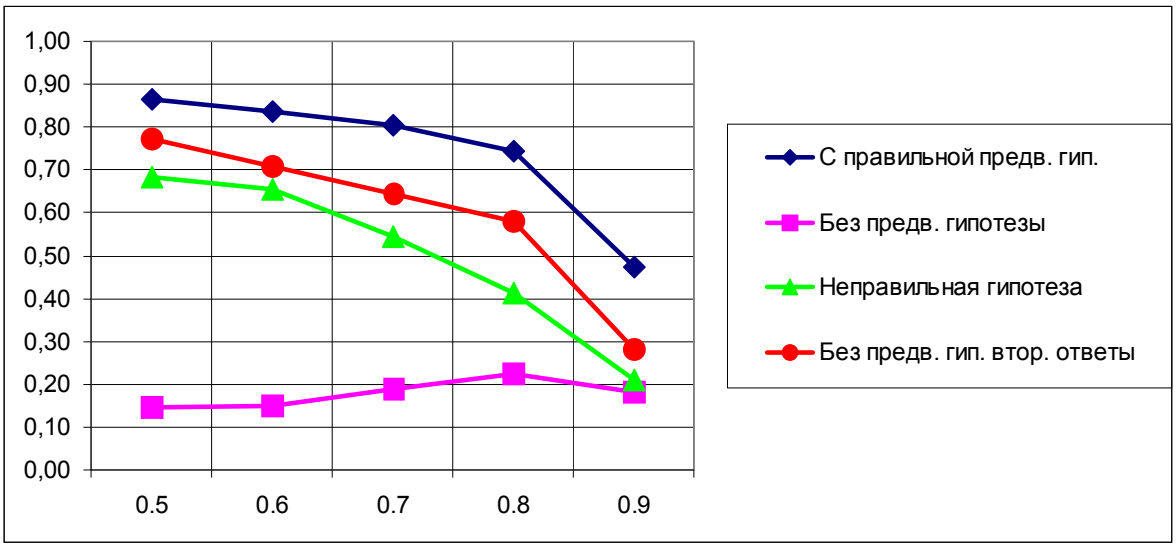


Рис. 7. Результаты исследования корректности алгоритма распознавания букв

Результаты показывают, что указание предварительной гипотезы позволяет добиться уровня распознавания до 87%. Отсутствие предварительной гипотезы резко снижает качество распознавания – 15-25%. При этом, в отсутствии предварительной гипотезы достаточно велик процент случаев, когда правильный ответ был выработан, но не сочтён наиболее правдоподобным – до 77% случаев. Наконец, распознавание под управлением заведомо неправильной гипотезы в значительном количестве случаев (до 68%) приводит к подтверждению данной ложной гипотезы, несмотря на то, что на изображении находится другая буква. Последний результат объясняется свойствами применённого метода трассировки.

На основе результатов исследования сделаны следующие выводы:

1. Полученные практические измерения временных затрат исследуемых алгоритмов не превышают установленных теоретических предельных оценок.
2. Установленный теоретически характер зависимостей показателей эффективности алгоритмов от содержания базы знаний подтверждается практическими результатами.
3. Наличие правильной предварительной гипотезы значительно повышает вероятность правильного распознавания букв.

4. Способ трассировки изображений нуждается в уточнении в рамках предложенной методики; возможно, требуется применение других методов анализа изображений.
5. Эксперименты по распознаванию слов с эмуляцией модуля распознавания букв полностью подтверждают корректность разработанного алгоритма в решении задачи распознавания скорописи.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

В процессе диссертационного исследования получены следующие результаты:
1. Впервые в качестве объекта исследования с точки зрения его компьютерного распознавания рассмотрена древнерусская скоропись XVIIв. Проведён анализ скорописных документов XVIIв. «Отводных книг Онежского крестного монастыря», выявлены особенности, имеющие значение для построения методики их распознавания.
2. Проведен анализ существующих методов и систем оптического распознавания текста с целью возможного применения их для распознавания скорописных документов XVIIв. Обоснована необходимость разработки нового подхода, способа, алгоритмов и программных систем для их распознавания.
3. Предложен новый структурный подход к распознаванию скорописных документов, основными принципами которого являются:

- реконструкция начертаний рукописных символов путём трассировки их изображений;
- использование экспертных палеографических знаний при обучении системы;
- управление распознаванием механизмом выдвижения и проверки гипотез в двухуровневом контексте «буква-слово»;
- интерактивность процесса распознавания.

4. Предложен способ структурного описания начертаний букв и слов и способ представления таких описаний в базе знаний системы распознавания, учитывающий особенности предложенного подхода к распознаванию.
5. Разработаны алгоритмы распознавания букв и слов скорописи, опирающиеся на предложенную схему построения баз знаний. Проведено теоретическое и практическое исследование их эффективности и корректности.
6. Установлены и подтверждёны характеристики вычислительной сложности алгоритмов. Установлены зависимости времени выполнения алгоритмов от характеристик содержимого баз знаний. Получены экспериментальные оценки среднего времени распознавания одной буквы (134 мс.) и страницы текста (53,6 с.). Получены экспериментальные оценки точности распознавания: 70-90% с указанием правильной предварительной гипотезы и 60-80% без указания предварительной гипотезы.

7. Предложена методика распознавания скорописи XVIIв., позволяющая разрабатывать системы автоматизированного распознавания скорописных документов.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. *Зеленцов И. А.* Выдвижение и проверка гипотез в системе распознавания древнерусской скорописи // Информационные технологии и письменное наследие: материалы междунар. науч. конф. Уфа - Ижевск, 2010. С. 99–101.
2. *Зеленцов И.А.* Информационная технология интеллектуальной обработки скорописных документов XVII в. // Системы, методы, техника и технологии обработки медиаконтента: материалы междунар. науч.-тех. молодёжной конф. М., 2011.
3. *Зеленцов И. А.* Метод распознавания древнерусской скорописи // Компьютерная графика и математическое моделирование (Visual Computing): сб. тез. и докл. науч. школы для молодых учёных. М., 2009. С. 116–131.
4. *Зеленцов И.А., Филиппович Ю.Н.* Распознавание букв и слов древнерусской скорописи XVII в. // Наука и образование: электронное научно-техническое издание. М., 2011. №12. URL: http://technomag.edu.ru/doc/296965.html (дата обращения: 22.12.2011).
5. *Зеленцов И.А., Филиппович Ю.Н.* Распознавание образов на основе структурных фреймовых описаний в скорописных текстах XVII в. // Наука и образование: электронное научно-техническое издание. М., 2011. №12. URL: http://technomag.edu.ru/doc/296744.html (дата обращения: 22.12.2011).
6. *Филиппович Ю.Н., Зеленцов И.А.* Распознавание скорописи XVII века // Проблемы полиграфии и издательского дела. М., 2011. №3, С. 87-97.
7. *Зеленцов И. А.* Учебно-практические занятия по распознаванию древнерусской скорописи // Печатные средства информации в современном обществе (к 80-летию МГУП); секция «Электронные средства информации в современном обществе»; сб. тез. докл. науч. межвузовской конф. преподавателей, аспирантов, молодых учёных и специалистов. М., 2010. С. 26-29.