# МЕТОД РАСПОЗНАВАНИЯ ДРЕВНЕРУССКОЙ СКОРОПИСИ

***И.А. Зеленцов***
***(МГТУ им. Н.Э. Баумана, г.Москва)***

*Работа посвящена вопросу автоматизации процесса получения текстов древних рукописей в электронном текстовом представлении. Приводится описание особенностей рассматриваемых рукописей и связанные с ними сложности автоматизированного распознавания. Получен вывод о целесообразности использования структурного подхода к распознаванию на основе экспертных знаний. Предложен двухуровневый принцип функционирования системы распознавания и способ представления знаний о структуре букв и слов с помощью фреймовых сетей. Описаны алгоритмы распознавания слов и букв, основанные на принципе выдвижения и проверки гипотез.*

## 1. ВВЕДЕНИЕ

В настоящее время исследователями русской письменности накоплено большое количество древнерусских рукописей различных временных периодов. Для обеспечения возможности компьютерного анализа и электронного переиздания этих документов требуется их перевод в электронный вид. Значительный объем задачи, а также весьма узкий круг специалистов, обладающих знаниями в сфере древнерусского языка, порождают необходимость в автоматизации данного процесса. В данной статье приводится описание системы, осуществляющей автоматизированный перевод растровых изображений рукописей в электронное текстовое представление.

## 2. ПРЕДМЕТНАЯ ОБЛАСТЬ

Сложность решения задачи компьютерного распознавания находится в сильной зависимости от особенностей графического представления текста. Для машинопечатного текста характерны следующие закономерности. Текст располагается в виде одной полосы на странице; расположение полос на всех страницах одинаковы. Полосы состоят из строк, расположенных одна под другой. Строки расположены на горизонталях страницы; левые и правые края строк выровнены (кроме начальных и конечных строк абзацев). Слова текста располагаются в строках, следуя друг за другом слева направо через пробельные промежутки. Буквы в словах следуют одна за другой и разделяются небольшими промежутками, т.е. не пересекаются. Буквы имеют четко определённое шрифтом начертание, одинаковое по всему тексту. В скорописных документах текст так же располагается в виде одной полосы на странице, однако положение полос на разных страницах варьируется. Полосы состоят из строк, расположенных одна под другой. Строки расположены на линиях, близких к горизонталям страницы, но чаще всего отклоняются от них, имеют изгибы. Чёткого выравнивания по краям нет, в частности из-за присутствия декоративных росчерков. Слова текста располагаются в строках, следуя друг за другом слева направо через пробельные промежутки; последние могут быть слабо выраженными из-за выступающих элементов букв других строк. Буквы в словах следуют одна за другой и могут либо разделяться небольшими промежутками, либо соединяться (собственными или специальными штрихами), либо иметь случайные пересечения, в том числе с буквами соседних строк. Буквы имеют начертание, определённое почерком создателя рукописи, и это начертание может широко варьироваться на протяжении текста. Элементы букв могут иметь искажения, вызванные неточностью движения руки (неровности, пропущенные участки), а также дополнительные декоративные росчерки. На рисунке 1 приводится изображение фрагмента рукописи, иллюстрирующее указанные особенности.

Таким образом главными особенностями рассматриваемых рукописей являются:

- Отсутствие ровных строк;
- Широкая варьируемость начертаний букв;
- Возможные дополнительные декоративные элементы в буквах;
- Соединение и пересечение букв.

На основе данных наблюдений можно сделать заключение, что большинство допущений относительно распознаваемого текста, принимаемых при распознавании машинной печати, не применимы к скорописным документам. Так, нельзя полагать, что буквы имеют в большой степени одинаковое начертание. Более того, одна и та же буква может иметь в разных точках текста различные декоративные штрихи, не входящие в основной набор формирующих букву элементов. Нельзя рассчитывать на возможность выделения отдельных букв в изображении с помощью поиска обособленных скоплений чёрных точек. Кроме того, расположение соседних букв может отличаться по вертикали, поэтому можно лишь приблизительно обозначать место поиска очередной буквы.

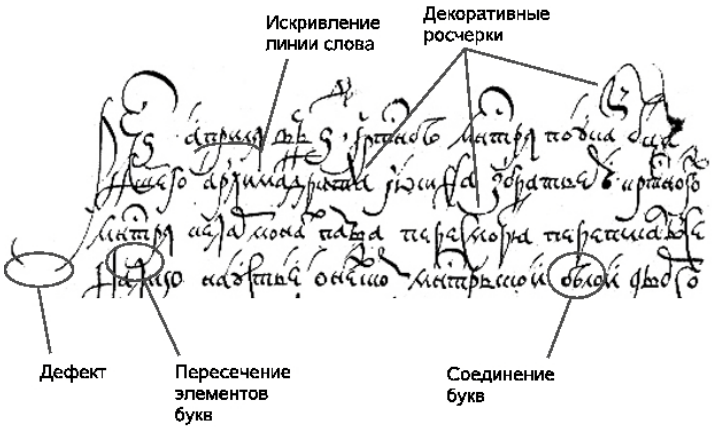


*Рисунок 1: Иллюстрация особенностей скорописного формирования*

## 3. ОСНОВНЫЕ ПРИНЦИПЫ

Как следует из описанных особенностей скорописи, система распознавания находится в ситуации, когда вариативность начертания распознаваемых символов велика, а их выделение из общего изображения затруднено. В связи с этим принимается структурный подход к распознаванию [5].

Для распознавания буквы необходимо выделить и определить её составные части. Этой цели может служить механизм векторизации входного растрового изображения. Специальный алгоритм (называемый *трассировщиком*) должен произвести анализ изображения и представить совокупности точек, образующие различные штрихи-элементы букв, в виде геометрических объектов, имеющих известные свойства. В решении этой задачи может быть задействован механизм скелетизации изображения [2,7] или метод восстановления траектории движения пера [3,4]. Другая часть системы (*распознаватель*), отвечающая за структурный анализ, должна выполнять оценку состава и отношений полученных примитивов. Наличие образцов структур букв позволяет распознавателю прогнозировать поступление информации об изображении и управлять продвижением трассировщика. Таким образом, задача выделения отдельных букв становится частью процесса распознавания.

Для преодоления трудностей распознавания, связанных с неточностью формирования элементов, в их описание вводится элемент нечёткости. Описание элемента отражает наиболее характерные его свойства, не привязываясь к конкретным измерениям его геометрии.

При оценке соответствия распознаваемого символа образцу единственное несоответствие может привести к отвержению гипотезы, если при этом пользоваться чётким понятием соответствия, имеющим два значения. Нечёткость на уровне анализа структуры позволит говорить и степени соответствия изображения образцу. Если эта степень превышает установленный в данной ситуации порог, то можно считать гипотезу о соответствии верной.

Процесс управления распознаванием основан на принятии гипотезы о наблюдаемом объекте и её целенаправленной проверке путём поиска предполагаемых элементов на изображении. Имея привязку к определённой точке изображения и предположения об окружающих её элементах, можно назначить последовательность проверок этих предположений, производя последовательный разбор изображения в соответствии с этим порядком. В случае неподтверждении гипотезы информация, полученная к данному моменту, сохраняется и служит для выбора другой гипотезы.

Подход, основанный на подтверждении гипотез, позволяет решить также и проблему непредсказуемых декоративных элементов букв и случайных пересечений, искажающих картину. Проверка гипотезы подразумевает поиск только тех элементов изображения, которые составляют образец предполагаемой буквы, и оставляет без внимания все лишние факты. Таким образом, гипотеза позволяет как-бы выделить суть из зашумлённой картины.

Для повышения вероятности правильного распознавания в метод включается использование словника, т.к. набор слов, используемых в рукописях, является в принципе известным. С появлением словника вводится ещё один промежуточный контекст — распознавание слов. Он позволяет усилить принцип выдвижения гипотез. Распознавая очередное слово, можно сделать предположение, какое из слов словника наблюдается в данный момент. Проверка такой гипотезы будет заключаться в последовательном распознавании всех букв слова, т.е. серии вызовов распознавателя букв (РБ) с указанием ожидаемой буквы. В случае, если гипотеза о букве не подтверждается, РБ возвращает распознавателю слов (РС) фактически обнаруженную букву, в результате чего последний корректирует свои прогнозы.

### 2.1 ЭКСПЕРТНЫЙ ПОДХОД

Все особенности древнерусского языка известны сегодня только кругу специалистов-исследователей. Таким образом, для обеспечения системы необходимыми данными о распознаваемом виде письма требуются знания эксперта, организованные в базу знаний. Первая её компонента содержит информацию о структуре изображения каждой буквы алфавита, причём буква может иметь несколько принципиально разных начертаний. Вторая компонента содержит информацию о структуре слов, отражающая как буквенный состав каждого слова, так и правила расположения изображений букв внутри изображения слова.

Кроме того, участие эксперта необходимо также в процессе работы системы. В случае неразрешимой ситуации система выдаёт эксперту описание проблемы, изображение проблемного участка рукописи и запрашивает правильный результат распознавания, а также область изображения, с которого следует продолжить процесс.

## 3. БАЗА ЗНАНИЙ

### 3.1 ФРЕЙМОВОЕ ПРЕДСТАВЛЕНИЕ ЗНАНИЙ

В качестве способа представления знаний предлагается использование фреймовых сетей, предложенные в [1, 6]. Сетевая природа фреймового представления позволяет корректно описывать сложный набор взаимосвязей структурных элементов букв. Кроме того, фреймовое описание хорошо согласуется с принципом проверки гипотез. Так, выдвижение высокоуровневой гипотезы означает активацию соответствующего фрейма, т.е. выделение из общей сети подмножества узлов, детализирующих его. Проверка заключается в попытке нахождения в распознаваемом изображении таких деталей, которые можно привязать ко всем терминальным узлам данной подсети.

### 3.2 СТРУКТУРА БАЗЫ ЗНАНИЙ

В соответствии с предлагаемым методом распознавания, в качестве ключевых понятий базы знаний следует выделить ``слово" и ``букву" как являющиеся объектами распознавания. База знаний представляется состоящей из множества фреймов, обозначающих конкретные слова, и множества фреймов, обозначающих буквы. Каждое слово состоит из некоторого количества букв, и все слова используют буквы из одного множества — алфавита языка. Таким образом, фрейм слова содержит набор субфреймов, описывающих вхождения составляющих его букв и связь между ними. Субфреймы-описатели слов ссылаются на фреймы-буквы как на детализирующие узлы, имеющие, в свою очередь, собственные субфреймы-описания. Описания букв устанавливают взаимоотношения между структурными элементами букв — линиями различных видов. Набор элементов также ограничен.

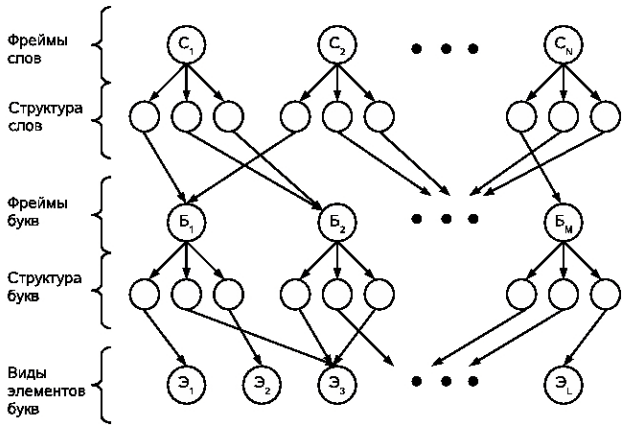


*Рисунок 2: Структурная схема системы распознавания*

Словник системы представляется в виде набора фреймов-слов, распознаваемый алфавит — в виде набора фреймов-букв. На рисунке 2 изображена общая картина содержания базы знаний. Из рисунка видно, что фреймы-слова разделяют фреймы-буквы, которые в свою очередь разделяют фреймы-элементы.

### 3.3 ОБЩАЯ СХЕМА ПОСТРОЕНИЯ ФРЕЙМОВЫХ МОДЕЛЕЙ

Описание детализируемого объекта строится из перечисления видов составляющих его элементов и их взаимоотношений. Причём, с одной стороны, объект может иметь несколько вхождений элементов одного вида, с другой стороны, различные объекты могут иметь одинаковые элементы, возможно имеющие одинаковые взаимоотношения. Здесь открывается возможность использования одного из свойств фреймовых сетей — разделения объектами частей описаний. Для этого следует разделить понятия элемента объекта и его вхождения в объект. С другой стороны, можно объединить понятие элемента с понятием отношения элементов на основе общего признака разделяемости, и назвать обобщённое понятие ``свойством" объекта. Тогда описание объекта будет состоять из набора *вхождений* различных *свойств*. Понятие вхождения отражает единичный факт присутствия в объекте того или иного свойства (элемента или отношения). И свойства составляющих частей объектов, и факты их вхождений в детализируемые объекты, представляются в фреймовой сети узлами специальных типов.

### 3.4 МОДЕЛЬ ПРЕДСТАВЛЕНИЯ БУКВ

Структурными элементами буквы являются линии определённых видов (элементарный штрих, входящий в начертание буквы), которые могут пересекаться. Полное описание структуры буквы можно построить, перечислив и охарактеризовав составляющие её линии, описав их пространственные

взаимоотношения и пересечения.

Любую линию характеризует её траектория, которую можно описать с помощью цепного кода направлений перемещения трассирующей точки при её движении от начала линии к концу [5]. Последовательность L измерений в градусах, разделённых символом ';', записывается в строку и называется в данной работе *путём* линии. При сравнении путей каждое чёткое угловое измерение в пути представляется в виде нечёткой величины, характеризующей примерное направление вектора.

Для учёта возможного различия масштабов линий с одинаковыми путями вводится ещё одна характеристика линии — форма. Эта характеристика является нечёткой лингвистической переменной и отражает общий вид линии — является линия более широкой или высокой. Вычисляется она по форме описывающего линию прямоугольника, а именно по отношению его высоты и ширины.

Пересечение двух линий можно описать, указав сами пересекающиеся линии и обозначив точку пересечения в каждой из них. Произвольную точку линии можно охарактеризовать положением по горизонтали и вертикали внутри описывающего прямоугольника линии. Нечёткое описание положения точки использует термины ``слева", ``в середине" и ``справа" для горизонтальной оси, и ``вверху", ``в середине" и ``внизу" — для вертикальной.

На рисунке 3 изображён упрощённый фрагмент фреймовой сети базы знаний. Здесь изображены возможные описания строчных букв 'н' и 'п'. В средней части рисунка располагаются общие для всех букв *Свойства*. Квадратами изображены *Линии*, а изображения внутри них заменяют описания самих линий. Кругами обозначены *Точки* пересечения линий, внутри кругов изображаются позиции точек. Ромбами показаны *Принадлежности Точек*, а шестиугольниками — *Соответствия Точек*. В верхней и нижней частях рисунка отображены *Вхождения Свойств* букв 'п' и 'н' соответственно. Белые треугольники обозначают *Вхождения Элементов*, а серые — *Вхождения Отношений*. Стрелками обозначены *отношения индикации* между *Вхождениями Свойств* и *Свойствами*. Простые соединительные линии обозначают *включения элементов* в отношения и *включения вхождений элементов* во вхождения отношений.

### 3.5 МОДЕЛЬ ПРЕДСТАВЛЕНИЯ СЛОВ

Фреймовые описания слов строятся аналогично описаниям букв. Свойствами слов являются буквы как элементы, и пространственные отношения между буквами. Отношения типа *Слева-Справа* определяют последовательность расположения изображений букв в изображении слова. Также используются отношения типа *Выше-Ниже*, что связано со спецификой древнерусской скорописи, в которой некоторые буквы в слове могут появляться над основным рядом букв.

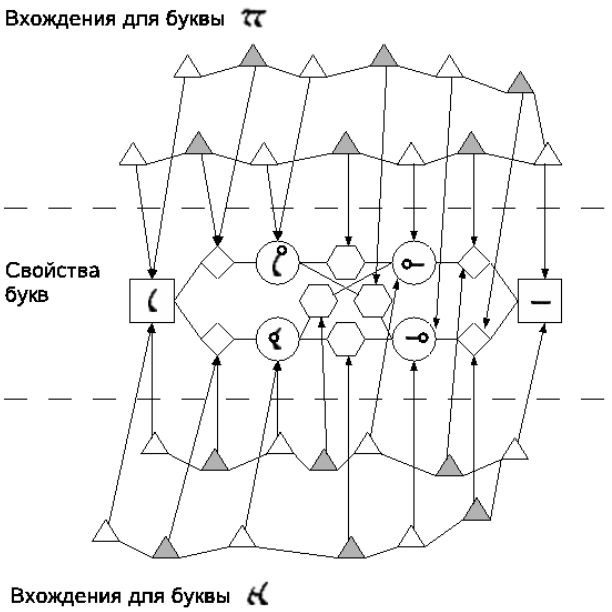


*Рисунок 3: Пример фреймовой сети, описывающей две буквы*

### 3.6 ОБУЧЕНИЕ СИСТЕМЫ

Источником знаний системы является эксперт. Для облегчения их ввода, система предоставляет специальный режим работы, называемый режимом обучения.

Для ввода в систему знаний о структуре букв эксперт в специальной области экрана рисует с помощью мыши или планшета каждую букву, выводя её линия за линией, тем самым определяя набор элементов. Система в онлайн-режиме осуществляет сбор информации о структуре буквы, классифицирует составляющие её линии, оценивает их взаимное расположение и размеры, определяет точки их пересечения и заносит полученную информацию в базу знаний.

Формирование словарной части БЗ осуществляется автоматически на основе имеющегося словника, представленного в виде списка слов.

## 4. РАСПОЗНАВАНИЕ

### 4.1 ВИРТУАЛЬНЫЙ ФРЕЙМ

В основе процесса согласования лежит концепция *виртуального фрейма* (ВФ). В процессе распознавания для сохранения получаемой информации об изображении в динамической памяти системы строится фреймовая модель, описывающая наблюдаемую в каждый текущий момент картину. Эта модель представляет собой не что иное, как описание буквы, аналогичное описаниям в БЗ. Отличие состоит только в том, что фреймы в БЗ статически описывают структуру эталонных начертаний букв, а в динамической памяти строится фрейм буквы,

рассматриваемой на изображении в данный момент, который дополняется новой информацией по мере её получения. Этот динамический фрейм и называется *виртуальным*.

### 4.2 ГИПОТЕЗЫ

Пусть в результате первого обращения к трассировщику получена линия некоторого типа, и в ВФ построено её описание, согласованное с БЗ. Можно предположить, что наблюдается одна из букв, содержащих в своей структуре линию найденного типа. Распознавание продолжается проверкой выдвинутых гипотез. Построенный динамический узел *Вхождения Элемента* может соответствовать такому же узлу одной из гипотез. Сами гипотезы представляются в виде специальных узлов, каждый из которых содержит ссылку на фрейм предполагаемой гипотезой буквы и набор ссылок на узлы согласований. Узел согласования содержит пару ссылок на согласованные узлы-вхождения из фрейма буквы и ВФ. На рисунке 4 изображен случай с двумя гипотезами.

Обозначим набор узлов *Вхождений Свойств* в фрейме-букве через Q, а набор таких узлов в ВФ — через V. Обозначим через N число пар *согласованных Вхождений* из Q и V. Тогда *степенью согласованности* гипотезы назовём величину

$$S_c = \frac{N}{|Q|},$$

а *степенью пригодности* гипотезы — величину

$$S_a = \frac{N}{|V|}.$$

Степень согласованности гипотезы отражает то, насколько полно она соответствует заданному фрейму. Эта величина будет расти по мере успешного согласования поступающей от сканера информации. При достижении некоторого верхнего порога $П_c$ можно говорить, что ВФ достаточно хорошо согласуется с данными фрейма и эта гипотеза может быть принята. Условие $S_c > П_c$ называется условием согласованности.

Степень пригодности говорит о точности соответствия ВФ рассматриваемому фрейму буквы. Чем меньше эта величина, тем больше в ВФ элементов, не имеющих отношения к букве-гипотезе. При успешном согласовании поступающей информации этот показатель должен оставаться вблизи максимального значения — 1. При снижении его значения ниже некоторого установленного порога $П_a$ гипотезу можно считать непригодной для дальнейшего рассмотрения. Будем называть условие $S_a > П_a$ условием пригодности.

После создания узлов для каждой из гипотез, для каждого из них вычисляются $S_a$ и $S_c$, и обладатель максимального значения $S_c$ принимается текущим. Дальнейшее распознавание ведётся путём проверки соответствующего ему фрейму из БЗ. При этом прочие гипотезы остаются в памяти. Более того, они участвуют в проверке параллельно с текущей. Выбор текущей гипотезы призван лишь определить наиболее перспективный путь распознавания.

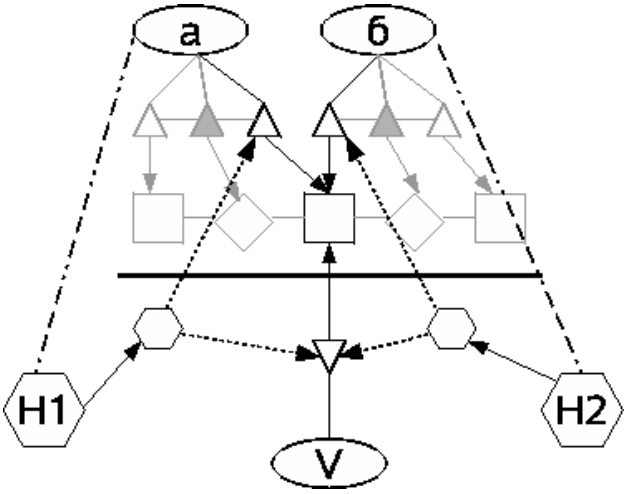


*Рисунок 4: ВФ и гипотезы*

Отталкиваясь от текущей линии, можно узнать из БЗ, какие точки пересечения должна иметь эта линия в наблюдаемой букве. Выполняется перебор этих точек, для каждой из которых выполняется следующее. Для текущей точки P определяется вид ожидаемой пересекающей линии и эта информация передаётся сканеру в виде гипотезы. Полученная от сканера линия заносится во ВФ. После этого определяются её реальные пространственные отношения с текущей линией и производится попытка её согласования с фреймами БЗ.

Далее происходит пересчёт $S_a$ и $S_c$ для всех гипотез. Нарушившие условие пригодности гипотезы исключаются из дальнейшего рассмотрения. Если нашёлся узел-гипотеза, удовлетворяющий условию согласованности, распознавание прекращается. Указанные действия повторяются для всех предполагаемых пересечений текущей линии. В конце концов ВФ будет содержать всю информацию о линиях, пересекающих текущую, а из числа гипотез будут исключены более не пригодные.

После согласования первой линии процесс дальнейшего распознавания разветвляется для выполнения согласования всех вновь найденных линий. Поочерёдно каждая из них принимается текущей во всех гипотезах и для неё выполняется та же процедура согласования. При этом каждый раз на основе значений показателей $S_c$ выбирается новая текущая гипотеза.

Условием окончания алгоритма является либо отвержение всех гипотез, либо выполнение условия согласованности для одной из них. В последнем случае буква считается распознанной. Вызывающему модулю РС возвращается узел *Вхождения* найденной буквы, *согласованный* с фреймом распознанной буквы в БЗ.

## СПИСОК ЛИТЕРАТУРЫ


1. Мински М. Фреймы для представления знаний. Пер. с англ. 1979.
2. Павлидис Т. Алгоритмы машинной графики и обработки изображений М.:Мир, 1982.
3. Поцепаев Р.Б., Петров И.Б. Эффективный алгоритм предобработки изображений для структурных методов распознавания рукописных символов [Электронный ресурс] / Моск. физ.-техн. ин-т. Электрон. журн. Долгопрудный : МФТИ, 2003.
4. Поцепаев Р.Б. Восстановление траекторий написания символов по их изображениям [Электронный ресурс] / Моск. физ.-техн. ин-т. Электрон. журн. Долгопрудный : МФТИ, 2003.
5. Фу К. Структурные методы в распознавании образов. М.: Мир, 1977.
6. Kuipers B. A Frame for Frames: Representing Knowledge for Recognition [Электронный ресурс] // Representation and understanding под ред. D. Bobrow and A. Collins, New York:Academic Press, 1975.
7. Крылов А.Б. Модуль предварительной векторизации растровых монохромных изображений гибридного редактора SpotLight // Интеллектуальные технологии и системы: сборник статей аспирантов и студентов / Под ред. Ю.Н.Филипповича — М.:МГУП, 2002. — Вып. 4.