Метод распознавания древнерусской скорописи

Целью разработки метода является решение задачи автоматизированного получения текстов древних рукописей в электронном текстовом представлении из растровых изображений.

Задачами исследования являются:

- изучение особенностей древнерусской скорописи;
- анализ существующих методов распознавания и разработка подходящего для решения задачи;
- проектирование и реализация системы распознавания.

Актуальность исследования обусловлена следующими обстоятельствами:

- необходимость представления рукописей в виде электронного текста для обеспечения их компьютерной обработки;
- ограниченность круга людей, способных к чтению данных рукописей, учёными-исследователями русского языка;
- трудоёмкость ручного перевода рукописей в электронное представление;
- отсутствие современных средств распознавания, работающих с древнерусской скорописью.

Скоропись имеет множество особенностей, затрудняющих её распознавание. Начертание букв сильно варьируется и 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.

Существует три основных вида методов распознавания текста. В методах, использующих Евклидово пространство, производится ряд измерений над матрицей изображения и совокупностью полученных значений задаётся точка в пространстве измерений. Классы распознаваемых объектов представляются в виде подпространств данного пространства. Признаковые методы выделяют в изображениях некоторые заданные характерные особенности, и объекты описываются как совокупности фактов наличия или отсутствия рассматриваемых признаков. Для этих методов характерна необходимость выделения распознаваемого объекта из общей картины. Это ограничивает их применимость к распознаванию скорописи, т.к. в данном виде текста отсутствуют явные границы букв и их выделение представляет собой отдельную сложную задачу.

Структурные методы основываются на выделении и анализе составных частей изображения. Структурные методы лишены недостатка других методов, т.к. при последовательном анализе элементов буквы можно постепенно и целенаправленно перемещаться по её изображению, попутно выявляя её границы.

Таким образом, в исследовании принят структурный подход к распознаванию. Для выявления элементов букв входное изображение подвергается векторизации, т.е. выделению на нём отдельных линий. Вариативность начертания букв преодолевается с помощью описания их структурных элементов и отношений между ними нечётким образом – на качественном уровне. Влияние случайных пересечений букв и декоративных росчерков исключается путём распознавания под управлением гипотез. Система выдвигает гипотезы о содержании наблюдаемого в данный момент фрагмента изображения и производит их проверку поиском предполагаемых ими элементов. Таким образом, выделяются только существенные части изображения, а дополнительные остаются без внимания.

В соответствии с разработанным методом, система распознавания представлена следующими компонентами: распознаватель, состоящий из распознавателя слов и распознавателя букв, трассировщик, база знаний и обучающий модуль. Основную функцию распознавания изображения выполняет модуль «распознаватель». Для распознавания текста

он выдаёт своему компоненту «распознаватель слов» серию команд на распознание всех слов текста. По каждой команде распознаватель слов выполняет серию вызовов компонента «распознаватель букв» для нахождения всех букв текущего слова. Для выполнения своей задачи распознаватель букв использует компонент «Трассировщик». Этот компонент выполняет векторизацию входного изображения рукописи и выделяет в ней отдельные линии, составляющие буквы. Распознаватель слов и распознаватель букв согласуют получаемые от нижележащих модулей данные с базой знаний, содержащей информацию о структуре изображений букв и слов. На сложных участках изображения, когда какой-либо из модулей оказывается не в состоянии выполнить свою задачу, выдаётся запрос к эксперту, управляющему работой системы. Он анализирует и разрешает сложившуюся ситуацию и позволяет системе продолжить работу далее.

Эксперт также осуществляет наполнение базы знаний системы путём диалога на высоком уровне с «Обучающим модулем». Он изображает вручную все буквы используемого алфавита, а система, наблюдая за процессом ввода, формирует в базе знаний описание структур букв. Информация о словах вводится в базу знаний путём анализа текстового файла словника. Таким образом реализуется экспертный подход к распознаванию.

В качестве метода представления знаний о структуре слов и букв выбраны фреймовые сети. Они позволяют эффективно описывать сложные структурные характеристики объектов и хорошо согласуются с методом проверки гипотез и нечётким описанием объектов.

Рассмотрим структуру используемой фреймовой модели описания распознаваемых объектов. Все фреймы в базе знаний имеют общий тип Узел. Основные узлы фреймовой сети имеют типы Слово и Буква. Они являются Детализируемыми узлами, т.к. все прочие узлы служат для описания их структуры. Их структурные элементы представляют узлы типа Элемент, а их взаимосвязи — узлы типа Отношение. Элементы и Отношения в общем называются Свойствами объектов.

Один объект, например буква, может иметь несколько однотипных Свойств. С другой стороны, одинаковые Свойства могут встречаться в разных объектах. Поэтому вводится различие между Свойством и его Вхождением в структуру конкретного объекта. Каждый узел типа Вхождение Свойства Детализирует какой-либо объект и Индицирует единичное присутствие в нём какого-либо Свойства. Различаются Вхождения Элементов и Вхождения Отношений. Таким образом, описание Детализируемого узла строится из набора *уникальных вхождений ограниченного набора Свойств*, разделяемых всеми объектами.

Элементами букв являются Линии и Точки. Принадлежность точек линиям описывается с помощью узлов-отношений типа принадлежность точки. Пересечения линий описываются с помощью Соответствий точек пересекающихся линий.

Конкретными видами Отношений являются Пространственные Отношения. Они предназначены для качественного описания относительных размеров и расположения любых объектов на изображении. Так, пространственное отношение типа Слева-Справа констатирует расположение одного объекта слева от другого с помощью слотов типа Слева и Справа. Аналогично вводятся отношения Выше-Ниже, Больше-Меньше и Равенство Размеров.

Отличительными характеристиками Линий разных видов являются нечёткие понятия Пути и Формы. Путь линии представляет собой набор измерений углов направлений обхода линии в фиксированных точках. При сравнении путей для точных измерений углов попарно вычисляются степени схожести и за результат принимается средняя степень схожести.

Форма линии определяется по её описывающему прямоугольнику. Чем больше угол его диагонали, тем с большей уверенностью его можно назвать высоким, а вписанную линию – вертикальной. Таким образом, сравнение форм описывающих прямоугольников ведётся в терминах Высокий, Широкий и, как промежуточный вариант, Квадратный.

Точки пересечения линий характеризуются их положениями внутри описывающих прямоугольников содержащих их линий. Положения описываются нечёткими множествами Слева, В Средней Части и Справа по горизонтали и Вверху, В Средней Части и Внизу по

вертикали. Эти характеристики вычисляются по отношению отстояний точки от левого верхнего угла прямоугольника к его измерениям по высоте и ширине.

На рисунке приведён пример описания двух букв. В средней части показаны Элементы и Отношения букв, в верхней и нижней – их вхождения в буквы П и Н. К примеру, буква П имеет вертикальную линию, которая пересекается в верхней части. Эта точка пересечения соответствует точке в левой части горизонтальной линии. Эта линия имеет ещё одну точку пересечения в правой части, которая соответствует верхней точке другой вертикальной линии. Описание буквы Н использует часть Свойств буквы П.

Элементами Слов, как сказано ранее, являются узлы-Буквы. Они связываются пространственными отношениями, описывающими расположение букв внутри изображения слова. На рисунке показано, как слова «сорокъ» и «горшковъ» используют общие узлы-Буквы. Эти узлы связаны отношениями Слева-Справа. Могут также присутствовать отношения Выше-Ниже, т.к. некоторые буквы имеют надстрочное написание.

Разбиение изображения на структурные элементы букв выполняет компонент системы, называемый “трассировщиком”. Перед началом распознавания трассировщик строит промежуточную внутреннюю модель изображения. Для этого применяется процедура истончения линий и строится граф тонких сегментов линий и точек их пересечения. При распознавании РБ обращается к сканеру с запросом на поиск очередной линии, указывая её параметры. От сканера требуется выполнить анализ графа изображения и выделить на нём группу узлов, представляющих линию искомого типа.

Процесс распознавания строится на основе концепции Виртуального фрейма. При распознавании база знаний является заданной и неизменяемой структурой. В процессе анализа изображения получаемая информация сохраняется в динамической памяти системы в виде фреймовой модели, описывающей наблюдаемую картину. Этот фрейм называется виртуальным. При обнаружении очередного элемента в виртуальный фрейм заносится узел, обозначающий Вхождение элемента данного типа в изображение буквы. Задача распознавания сводится к нахождению способа установления соответствия между узлами виртуального фрейма и узлами одного из фреймов букв в базе знаний. Этот процесс называется согласованием.

Выделив на изображении единственный элемент, можно построить первый список гипотез. Ими будут являться буквы, содержащие хотя бы один элемент найденного типа. С нахождением новых элементов на изображении этот список будет уменьшаться, т.к. всё меньше фреймов в базе знаний будут содержать полный набор найденных элементов.

Для описания выдвинутых гипотез для каждой из них в динамической памяти строится специальная структура, состоящая из пар ссылок на согласованные узлы и тем самым описывающая схему согласования. Структуры-гипотезы могут указывать и способы согласования информации с различными фреймами, и различные способы согласования с одним фреймом.

Для определения правдоподобности гипотез используются следующие характеристики. Степень согласованности гипотезы отражает, насколько полно в данный момент ВФ соответствует предполагаемому фрейму и определяется как отношение числа пар согласованных узлов в гипотезе к общему числу узлов в фрейме буквы. При успешном согласовании очередного узла она увеличивается, и как только она поднимется выше определённого порога, гипотезу можно считать подтверждённой. Степень пригодности гипотезы говорит о точности установленного соответствия и вычисляется как отношение числа пар в гипотезе к числу узлов ВФ. Чем больше в ВФ узлов, не согласованных данной гипотезой, тем ниже эта характеристика, и при определённом минимальном значении можно говорить о несостоятельности гипотезы.

При проверке гипотез для каждой из них вычисляются степени согласованности и пригодности и одна из них принимается текущей. С её помощью в базе знаний определяются ожидаемые пересечения текущей линии и типы пересекающих её линий. Далее трассировщику передаются запросы на поиск ожидаемых линий на изображении,

полученные результаты анализируются и заносятся в ВФ в виде Вхождений Элементов и Отношений. При этом проводится попытка согласования добавляемых узлов с фреймами базы знаний во всех гипотезах. Далее выполняется пересчёт характеристик гипотез и нарушившие условие пригодности удаляются. Если находится гипотеза, удовлетворяющая условию согласования, то считается, что фрагмент изображения распознан и ответом является буква, указываемая данной гипотезой. Если такой гипотезы нет, текущей назначается гипотеза с максимальной степенью согласованности и распознавание продолжается.

Распознавание слов во многом аналогично распознаванию букв. Вместо линий и точек здесь проводится согласование узлов букв и пространственных отношений между ними.

В качестве средства реализации выбран язык Java. Для описания знаний по предложенной схеме использован язык веб-онтологий OWL. Он имеет достаточный набор средств для описания всех видов объектов базы знаний и их взаимоотношений. Программная поддержка этого языка осуществляется библиотекой Jena. Она предоставляет средства для программного оперирования сущностями онтологий, а также организует преобразование OWL-моделей в реляционные структуры для хранения в базе данных.