Информационная технология интеллектуальной обработки скорописных документов XVII в.[*]

Зеленцов И.А., МГТУ им. Н.Э. Баумана, аспирант

В настоящее время исследователями русской письменности 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, обладающих знаниями в сфере древнерусского языка, порождают необходимость в автоматизации данных процессов.

Сложность автоматизации получения электронных текстовых представлений скорописных рукописей обусловлена спецификой используемого в них языка и стиля письма, а также их ветхостью. Кроме того, круг носителей знаний о лексике русского языка XVII в. и способах письма, использовавшихся в этот период, ограничен в настоящий момент немногочисленной группой специалистов в области палеографии, исторического языкознания и филологии. 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 распознавания, учитывающей особенности скорописного способа формирования текста, применявшегося в России XVII в.

Предложены следующие ключевые особенности информационной технологии распознавания. 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.

Разработан способ структурного описания изображений букв на основе реконструкции их начертаний в виде набора траекторий движения пера автора текста. Разработан способ представления таких структурных описаний в базе знаний системы распознавания на основе фреймовых сетей.

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.

Установлены и подтверждены 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.

[*] Работы поддерживаются грантом Президента РФ МК-3732.2010.9 «Разработка словарных компонентов интегрированной информационной технологии переиздания печатных источников XVIII – нач. XIX вв.».