МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

Писменото наследство и информационните технологии [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.

ISBN 978–954–9787–25–2

Модель системы распознавания старопечатных кириллических текстов с помощью лингвистической базы данных

**Д. Р. Касимов, А. В. Кучуганов,
М. Н. Мокроусов, П. П. Осколков**

Распознавание, скелетон, структурные элементы, нечеткий граф, грамматический словарь, нечеткий поиск

A Model for a System of OCR of Early Printed Cyrillic Texts with the Help of a Linguistic Database

**Denis Kasimov, Aleksander Kuchuganov,
Maksim Mokrousov, Pavel Oskolkov**

In this paper a technique and an experimental system of optical character recognition of early printed Cyrillic texts are proposed, which are characterized by building an attributed graph of an image containing fuzzy attributes of the outline and skeleton of symbols generated by detection of letters on the basis of description logic as well as by the method of fuzzy recognition of words by building a tree of correspondences to a linguistic database.

Актуальность задачи автоматизации процесса перевода древних кириллических рукописей из графического представления в текстовую форму обусловливается тем, что они представляют исключительную ценность для лингвистических исследований, проведение которых наиболее эффективно с использованием методов автоматического анализа текста, который невозможен, если рукопись доступна лишь в электронно-графическом виде.

Для максимальной автоматизации процесса оцифровки старинных текстов требуется привлечение средств распознавания образов. Относительно небольшое количество публикаций по системам распознавания рукописных и старопечатных кириллических текстов XI–XVIII веков говорит о необходимости совершенствования методов и технологий решения этой проблемы. Существующие подходы к распознаванию старославянских и древнерусских текстов имеют неудовлетворительную надежность распознавания: в среднем 60–70 % на изображениях неплохого качества.

В работе [Корниенко 2011] задача распознавания решается с помощью искусственных нейронных сетей. Подход требует, чтобы входные изображения имели очень высокое разрешение.

В работе [Зеленцов, Филиппович 2011] предлагается подход к распознаванию, ориентированный на скорописные тексты. Осуществляется выделение, нечеткое фреймовое описание и сравнение структурных элементов. При этом реализуется

двухуровневое распознавание (слово-буква), управляемое гипотезами, с применением словарной информации. Слабым местом подхода, на наш взгляд, является используемый метод трассировки, а также недостаточная информативность описания структурных элементов.

В настоящей работе предлагается новая методика анализа и распознавания древнерусских текстов, включающая 6 этапов:

- классификация букв по степени сходства и надежности распознавания;
- цветовая сегментация;
- выделение и аппроксимация границ областей и их скелетонов;
- фаззификация и формирование нечеткого атрибутивного графа изображения;
- выделение элементов букв, вспомогательных и декоративных знаков с помощью дескрипционной логики ALC;
- нечеткое распознавание слов и построение дерева соответствий лингвистической базе данных.

Выделенная на изображении буква представляется нечетким пространственно-нагруженным графом структурных элементов, в котором все атрибуты нечеткие.

На рисунке 1 представлены типовые структурные элементы букв.

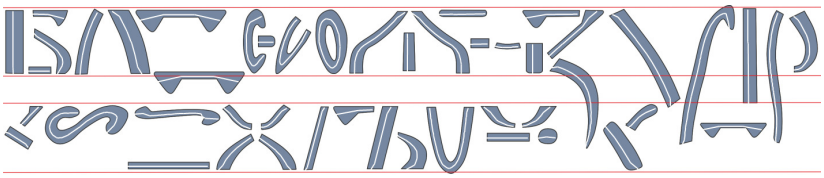


Рис. 1. Типовые структурные элементы букв

Атрибуты структурного элемента: форма (прямая, левый плавный поворот, левый крутой поворот, правый плавный поворот, правый крутой поворот, извилистая), длина, ширина начала, ширина середины, ширина конца, ориентация.

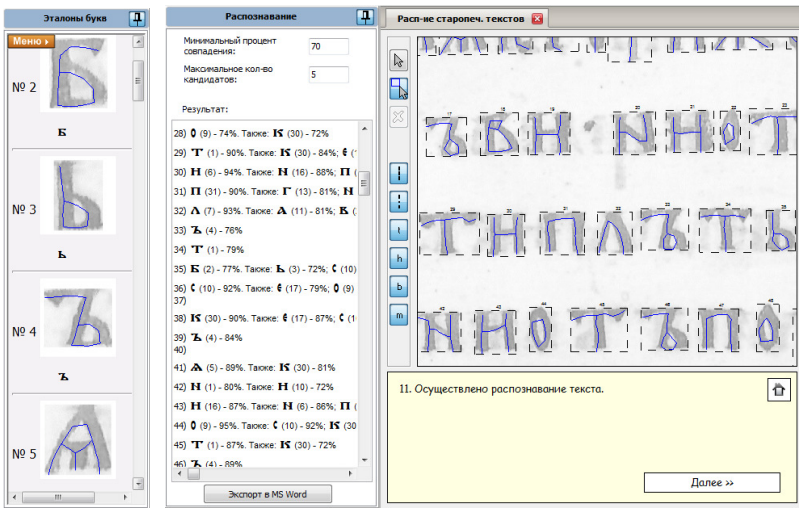
Отношения между структурными элементами: касание (концевая точка с концевой, концевая точка с промежуточной, промежуточная точка с промежуточной), пересечение, положение (близко и сверху, далеко и сверху-слева и т. д.).

Распознавание букв осуществляется на основе сравнения с эталонами. Эталоны должны быть построены на высококачественных изображениях букв. Эталон буквы включает растр, границы и скелетон, граф структурных элементов. На рисунке 2а представлен фрагмент библиотеки эталонов.

В экспериментах использовался текст Остромирова Евангелия 1056–1057 гг. (294 л., РНБ, F.п.1.5.). По одной странице указанного текста в систему было заведено 38 эталонов. На другой распознаваемой странице присутствовало 272 символа, из которых 209 (77 %) были распознаны правильно и еще 32 символа были

распознаны неоднозначно, т. е. процент потенциально распознанных составляет 89 %. На рисунке 2б представлен пример результатов распознавания текста.

Полученный в результате распознавания текстовый документ содержит последовательность наборов претендентов на итоговый результат распознавания отдельных символов текста. Претенденты символа указаны в квадратных скобках в порядке убывания их релевантности. С целью повышения надежности распознавания символов использовался словарь древнерусского языка объемом около 2 млн. словоформ. Словарь создан в рамках проекта “Манускрипт” (manuscripts.ru) и предоставлен коллективом этого проекта.



а) б)

Рис. 2. Модуль распознавания:
а) библиотека эталонов; б) результаты распознавания

В ходе работы алгоритма поиска строятся все возможные варианты предложений из кандидатов букв и с помощью запросов полнотекстового поиска вычисляются совпадения слов, начиная с самого короткого длиной 3, заканчивая наиболее длинным, для которых возвращается наименьшее количество совпадений. Сочетания букв, для которых возвращается одно совпадение, принимаются за константы, а поиск повторяется дальше, начиная со следующих символов. Если для каких-то вариантов сочетаний не возвращается никаких данных по запросу, то в таком случае строятся регулярные выражения, учитывающие в своей конструкции буквы, имеющие наибольшую релевантность по результатам распознавания.

В ходе эксперимента (рис. 3) на фрагменте текста из 21 буквы, каждая из которых содержала от 1 до 4 кандидатов, в результате полнотекстового поиска по грамматическому словарю древнерусских словоформ было проведено сокращение кандидатов: из 65 рассматриваемых кандидатов было удалено 38, а у 18 букв из всех кандидатов осталось по 1. Таким образом, процент надежности распознавания букв был увеличен до 85 %.

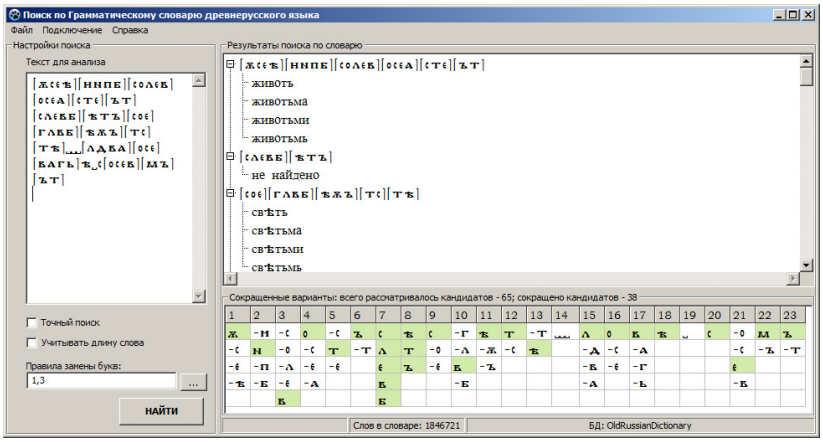


Рис. 3. Пример подпрограммы сокращения вариантов распознавания с помощью словаря древнерусских слов

Несомненно, итоговый текст должен быть проверен на синтаксическую и семантическую корректность, т. к. в ходе поиска по словарю учитываются лишь буквенные последовательности, которые с наибольшей долей вероятности могут являться словами языка.

Следует отметить, что применение грамматического словаря древнерусских словоформ будет целесообразно при выполнении двух условий:

- при выполнении запросов необходимо учитывать морфологические правила преобразований букв и их сочетаний;
- словарь должен быть расширен словами с сокращениями и их расшифровками.

Таким образом, применение грамматического словаря позволит решить две задачи: во-первых, сократить количество кандидатов распознавания букв, т. е. повысить точность распознавания, и, во-вторых, дать оценку эффективности предлагаемой методики и программной системы графического распознавания древнерусских рукописей.

Литература

Зеленцов, Филиппович 2011 — *Зеленцов И. А., Филиппович Ю. Н.* Распознавание образов на основе структурных фреймовых описаний в скорописных текстах XVII в. [Электронный ресурс] // Наука и образование: электронное научно-техническое издание. М., 2011. № 12. Режим доступа: http://technomag.edu.ru/doc/296744.html, свободный (дата обращения: 13.02.2014).

Корниенко и др. 2011 — *Корниенко С. И.* Программный комплекс для распознавания рукописных и старопечатных текстов / С. И. Корниенко, Ю. Р. Айдаров, Д. А. Гагарина, Ф. М. Черепанов, Л. Н. Ясницкий // Информационные ресурсы России. 2011. № 1. С. 35–37.