МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. М. Т. КАЛАШНИКОВА

Информационные технологии и письменное наследие

El'Manuscript–2012

Материалы IV международной научной конференции
Петрозаводск, 3–8 сентября 2012 года

Петрозаводск, Ижевск
2012

УДК 004.9
ББК 81.11+81.2-0
И741

Изданы при поддержке гранта РФФИ (проект № 12–06–06061–г),
гранта РГНФ (проект № 12–04–14154–г) и в рамках реализации комплекса
мероприятий Программы стратегического развития ПетрГУ на 2012-2016 г.

Ответственные редакторы:
В. А. Баранов, д-р филол. наук, проф.
А. Г. Варфоломеев, канд. физ.-мат. наук, доц.

И741 **Информационные технологии и письменное наследие** [Текст] : материалы IV междунар. науч. конф. (Петрозаводск, 3–8 сентября 2012 г.) / отв. ред. В. А. Баранов, А. Г. Варфоломеев. — Петрозаводск ; Ижевск, 2012. — 328 с.

ISBN 978–5–8021–1402–5

Сборник содержит материалы конференции, посвященной современным электронным средствам хранения, описания, обработки, исследования и публикации памятников письменности и исторических источников.

УДК 004.9
ББК 81.11+81.2-0

ISBN 978–5–8021–1402–5

ЭКСПЕРИМЕНТЫ ПО РАСПОЗНАВАНИЮ БУКВ И СЛОВ СКОРОПИСИ XVII ВЕКА

И. А. Зеленцов

«ТрансИнфоСеть», Москва

In previous works a new approach to recognition of XVIIth century Russian handwritten documents was proposed. In this paper we describe experiments that were carried out with the developed pilot recognition program to figure out the correctness of the proposed approach.

Контекст исследований

Автором доклада предложена методика распознавания древнерусских скорописных документов XVII в. Она характеризуется применением струк-турного подхода к распознаванию, использованием экспертных палеографи-ческих знаний в виде фреймовых сетей, распознаванием под управлением гипотез в двухуровневом контексте «буква-слово». В рамках исследования методики был реализован опытный программный комплекс, реализующий модули распознавания букв (РБ), распознавания слов (РС) (Java), модуль обучения и управления базами знаний (БЗ) (Java, OWL), модуль трассировки на основе истончения линий изображения (C++). В настоящей работе описа-ны эксперименты по распознаванию изображений отдельных букв скорописи с использованием полученного программного комплекса.

Методика экспериментального исследования

Для проведения исследования корректности алгоритма распознавания букв был сформирован набор баз знаний, содержащих структурную инфор-мацию о 60 начертаниях 16 букв алфавита. Для формирования каждой из БЗ был использован один и тот же набор начертаний символов. Различие за-ключалось в том, какой порог нечёткого сравнения $M_{об}$ линий применялся при создании БЗ. В зависимости от величины порога при обучении в БЗ формировались различные наборы линий E_{LN}, выступающих в качестве структурных элементов начертаний букв. Чем выше этот порог, тем ниже ве-роятность признания двух линий одинаковыми и, следовательно, тем больше типов линий составляют элементную базу описаний. В таблице 1 приводятся количественные характеристики полученных баз знаний.

Табл. 1. Характеристики баз знаний, использованных для проверки кор-ректности алгоритма распознавания букв

№ пп	Название характеристики	Обозн.	БЗ 1	БЗ 2	БЗ 3	БЗ 4	БЗ 5
1	Порог сравнения обучения для линий	$M_{об}$	0,5	0,6	0,7	0,8	0,9
2	Число типов линий	$\|E_{LN}\|$	3	4	6	12	28
3	Число начертаний букв	$\|D_L\|$	60	60	60	60	60
4	Среднее число вхождений линий в начертании буквы	$\tilde{n}_{БЛ}$	2,5	2,5	2,5	2,5	2,5
5	Среднее число вхождений линий различных типов в начертании буквы	$\tilde{n}_{БР}$	1,72	1,82	1,9	2	2,08
6	Среднее число вхождений линии данного типа во все начертания	$\tilde{n}_{ЛБ}$	50	37,5	25	12,5	5,35
7	Расчетное число обращений к трассировщику для средней буквы	$\tilde{c}_{\max}$	125	93,75	62,5	31,25	13,375

ных букв скорописи (рис. 1). Данные изображения были получены путём ручного выделения начертаний букв из сканированных изображений скоропи-сных документов.

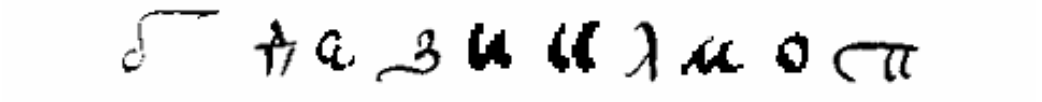

Рис. 1. Примеры распознаваемых изображений отдельных букв

Суть пакетного эксперимента по распознаванию букв заключалась в следующем. Для каждой из баз знаний выполнялась серия распознаваний изображений отдельных букв, начертания которых описаны в БЗ, с формированием отчёта о полученных результатах. Распознавателю букв предъявлялось по 400 различных изображений каждой буквы. В качестве показателя корректности распознавания был принят процент правильно распознанных букв, т.е. отношение количества циклов распознавания, в которых в качестве ответа была названа буква, изображённая на распознаваемом изображении, к общему числу циклов распознавания.

Исследование состояло из трёх групп экспериментов. Каждая из групп состояла из пакетных экспериментов над всеми базами знаний с разными значениями порога сравнения распознавания $M_{расп}$ для линий: от 0,5 до 0,9 с шагом 0,1. Т.е. в каждой группе экспериментов проводилось 25 экспериментов — по числу комбинаций значений порогов распознавания и обучения.

Первая группа экспериментов проводилась с указанием модулю РБ предварительных гипотез, заведомо соответствующих распознаваемым изображениям. Во второй группе экспериментов указание какой-либо предварительной гипотезы не производилось. При этом измерялась доля правильных выданных ответов, а также доля правильных ответов, которые содержались в списке второстепенных ответов, но не были выданы в качестве основных ответов распознавания. Третья группа экспериментов была направлена на получение доли успешных распознаваний при условии, что подавалась заведомо неверная предварительная гипотеза. В таком случае, очевидно, процент правильных ответов равен 0, т.к. при указании предварительной гипотезы прочие варианты не ответов не рассматриваются. Задачей же третьей группы экспериментов было выяснить, насколько возможна выдача неверного ответа модулем РБ под влиянием заведомо ложной предварительной гипотезы.

Результаты исследований

На рисунке 2 представлены графики результатов экспериментов при $M_{об}= M_{расп}$. Анализ результатов позволяет сделать следующие наблюдения:

1. Указание предварительной гипотезы позволяет добиться уровня распознавания до 87%. 2. Отсутствие предварительной гипотезы резко снижает качество распознавания — 15–25%. 3. При этом в отсутствии предварительной гипотезы достаточно велик процент случаев, когда правильный ответ был выработан, но не сочтён наиболее правдоподобным — до 77% случаев. 4. 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.

Наблюдения 1, 2 и 3 являются в той или иной мере закономерными. Наличие правильно предварительной гипотезы значительно повышает корректность распознавания, что и являлось целью введения данного механизма в процесс распознавания.

Наблюдение 4 в то же время носит отрицательный характер. Вышестоящий модуль РС, проводя распознавание слова, выполняет серию проверок букв, чтобы иметь возможность определения правдоподобности выдвинутых им гипотез. Возможность подтверждения модулем РБ заведомо неверных гипотез приводит к тому, что неверные гипотезы относительно слов будут получать поддержку, а верная гипотеза (если она присутствует) будет опровергаться, что ведёт к выдаче неверного ответа в распознавании слова либо к невозможности выдачи какого-либо ответа.

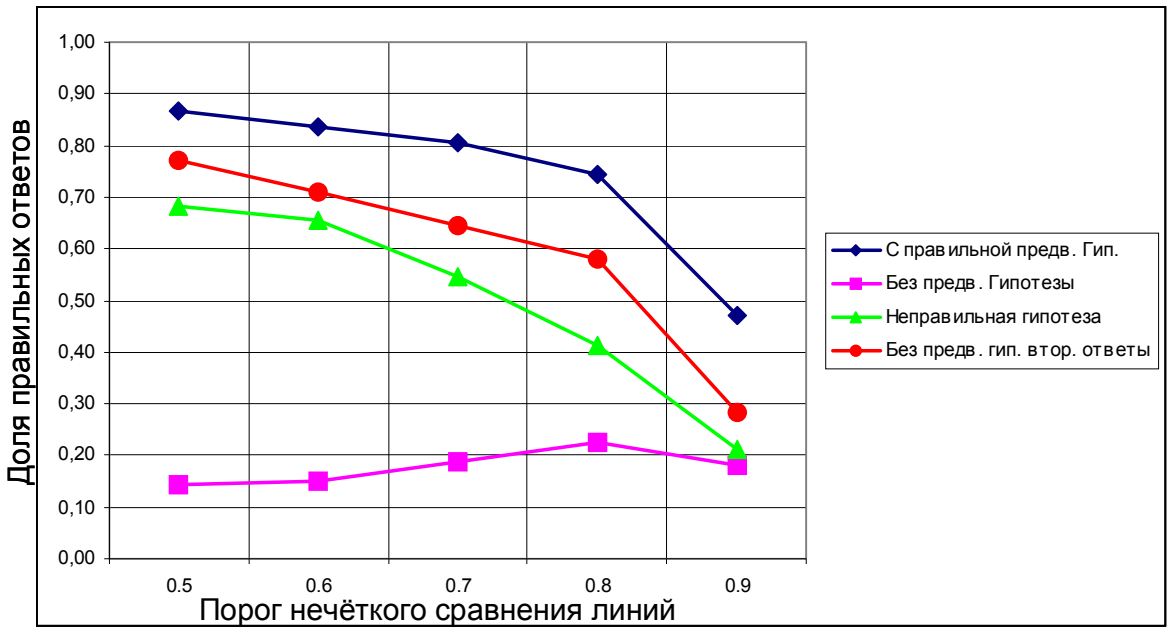

Рис. 2. Результаты проверки корректности алгоритма распознавания букв

Это негативное свойство распознавания букв объясняется недостатками использованного способа трассировки изображения на основе процедуры истончения линий. Основным недостатком этого способа является возможность множественной интерпретации данных. Модель изображения в виде графа тонких сегментов линий позволяет по-разному компоновать узлы этого графа в процессе поиска линий с указанными свойствами. Это ведёт к тому, что одна и та же часть изображения буквы может быть рассмотрена как компонент более чем одного структурного состава начертания. И в зависимости от параметров запроса эта часть изображения может включаться в состав различных структурных элементов буквы. Таким образом, неверная гипотеза может привести к истолкованию изображения соответственно предположениям этой гипотезы, что является ошибочным действием.

Замечания об эффективности распознавания

В работах автора сформулированы алгоритмы распознавания букв и слов, а также выведены выражения, дающие оценку максимального числа узлов, которые необходимо проверить в ходе выполнения алгоритмом про-цедуры выдвижения и проверки гипотез относительно наблюдаемой буквы или слова. Иными словами, это выражение показывает, сколько в худшем случае потребуется обращений к модулю трассировки для распознавания бу-квы или к модулю РБ для распознавания слова, имеющих средние по базе знаний структурные описания.

Проведённые в исследовании эксперименты позволили практически измерить указанные характеристики эффективности алгоритмов. Получен-ные результаты показали соблюдение характера теоретических зависимостей и непревышение среднемаксимальных оценок. Среднее время распознавания буквы по всем экспериментам составило 134 мс. При средней оценке числа букв на странице скорописного документа в 400 символов можно ожидать среднее время распознавания страницы равным 53,6 с.

Заключение

На основе полученных наблюдений сделаны следующие выводы:

1. Важным с точки зрения корректности распознавания является нали-чие предварительной гипотезы. При построении системы распознавания сле-дует стремиться к минимизации распознаваний без предварительной гипоте-зы. Для модуля РБ источником предварительных гипотез является РС. Для модуля РС, в свою очередь, источником предварительных гипотез могут яв-ляться другие, более высокоуровневые модули синтаксического и семанти-ческого анализа, а также независимые модули, например, предварительного синтагматического анализа подъязыка класса рукописей, основной функцией которых было бы получение основных статистик встречаемости различных комбинаций букв и слов. Предлагаемая архитектура построения системы распознавания открыта для дополнения числа уровней распознавания, что позволяет добавлять к предложенной двухуровневой схеме вышележащие модули в зависимости от потребностей и результатов распознавания. Про-дуктивным также видится применение квантитативных лингвистических данных уровня слов и словосочетаний.

2. Критическим является использование модуля трассировки, отвечаю-щего требованию единственности интерпретации данных изображения. Представляется, однако, маловероятной возможность реализации трасси-ровщика, полностью исключающего множественность интерпретации. В та-кой ситуации следует стремиться к минимизации этого качества за счёт, на-пример, более интенсивного использования морфологической информации в распознаваемых изображениях.

Список литературы

Зеленцов, 2010а — 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.

Зеленцов, 2010б — Зеленцов И.А. Учебно-практические занятия по рас-познаванию 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.

Зеленцов и Филиппович, 2011а — Зеленцов И.А., Филиппович Ю.Н. Распознавание букв и слов древнерусской скорописи XVII в. // Наука и обра-зование: электронное научно-техническое издание. М., 2011. № 12. URL: http://technomag.edu.ru/doc/296965.html (дата обращения: 22.12.2011).

Зеленцов и Филиппович, 2011б — Зеленцов И.А., Филиппович Ю.Н. Распознавание образов на основе структурных фреймовых описаний в ско-рописных текстах XVII в. // Наука и образование: электронное научно-техническое издание. М., 2011. № 12. URL: http://techno-mag.edu.ru/doc/ 296744.html (дата обращения: 22.12.2011).

Филиппович и Зеленцов, 2011 — Филиппович Ю.Н., Зеленцов И.А. Распознавание скорописи XVII века // Проблемы полиграфии и издательско-го дела. М., 2011. № 3, С. 87–97.