

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.

Информационные технологии и письменное наследие [Текст] :
И741 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 современ-
ным 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

© Петрозаводский государственный
университет, 2012
© Ижевский государственный технический
университет им. М. Т. Калашникова, 2012

О ДЕШИФРОВКЕ РУКОПИСНЫХ ИСТОРИЧЕСКИХ ДОКУМЕНТОВ

А. А. Рогов, А. В. Скабин, И. А. Штеркель

Петрозаводский государственный университет, Петрозаводск

Article contains research results of creation of a program system for automated recognition of historical handwritten documents, including historical shorthand reports of XIX and 1st part of XX centuries. A problem of receiving original images of symbols from historical documents using threshold binarization and search for similar symbols in a knowledge base is considered. In addition, a description of the prototype of the automated program system for deciphering historical handwritten documents is presented.

Одним из важнейших вопросов истории является проблема понимания текстов. На сегодняшний день в архивах России накопился большой объем нерасшифрованных стенографических документов. В XIX и начале XX веков стенография в России находилась в процессе становления, поэтому существующие документы записаны в разных системах. Основной проблемой при дешифровке исторических стенограмм является отсутствие специалистов, обладающих знаниями о системах стенографической записи того времени. Еще одной трудностью дешифровки является то, что стенографист при записи мог использовать свои нестандартные символы (обозначения), так как часто расшифровкой занимался он сам.

Таким образом, для введения в научный оборот новых документов необходимо описание и дешифровка исторических стенограмм. Для решения этой задачи необходимо создать универсальную программную систему автоматизированного распознавания исторических рукописных текстов, включая исторические стенограммы XIX и начала XX веков. К отличительным свойствам разрабатываемой системы относятся: учет индивидуальных знаков разных стенографистов, возможность критического анализа, использование словаря для подсказки при дешифровке текста и т. д. [Рогов и др., 2011]. Информационная система будет находиться в открытом доступе и предлагаться к использованию работниками архивов, научными сотрудниками, исследователями текстологам. Было принято решение отлаживать систему на стенограммах А.Г. Сниткиной, частично расшифрованных Ц.М. Пошемянской, и учебнике П. Ольхина [Ольхин, 1866]. Данное исследование поддержано грантом РГНФ № 11–01–12026в (рук. Рогов А.А.).

Для распознавания стенограмм необходимы оригинальные символы, который использовал стенографист в своих записях, так как стенографист мог видоизменять и модернизировать символы стенографической системы, которую он использовал. Поэтому было написано приложение, позволяющее получать оригинальную графику символов из стенограмм. Для выделения ори-

гинальных символов и последующей работы с ними изображение стенографического документа подвергалось процессу бинаризации.

Выделение символов на рукописных исторических документах имеет следующие особенности:

- оригинальное изображение имеет низкое качество из-за того, что рукописям порядка 150 лет, запись производилась простым карандашом на пожелтевшей бумаге, кроме того, на рукописях есть посторонние записи, пометки, которые не несут смысловую нагрузку, но пересекаются со стенографическими записями.

- при бинаризации происходят разрывы символов, т.к. некоторые пиксели символа имеют схожий цвет с пикселями фона;

- при сегментации возникает необходимость разбиения символов, написанных слитно, на отдельные символы.

Было обработано 29 листов стенографических записей и получено более 2500 изображений символов. После этого возникла задача поиска похожих символов и их кластеризации.

Поиск с использованием эталона заключается в том, что изображение символа используется как шаблон и сравнивается с изображениями других символов. При сравнении изображений вычисляется расстояние Хэмминга между двумя матрицами из нулей и единиц.

Данный метод обладает высокой точностью поиска при пороговых значениях уровня совпадения от 80%, но при этом полнота поиска крайне мала. Логическое сравнение с эталоном показывает плохие результаты для сложных символов, что происходит из-за того, что человек пишет символы по-разному. Меняются углы, размер и толщина символа. При сравнении скелетов символов исключается различие толщины символов, но усиливается различие в зависимости от других факторов написания символа. Скелетизация символов производилась алгоритмом Зонга-Суня [Zhang and Suen, 1984].

В ходе исследования был проведен анализ метода краевых расстояний. Принцип его работы заключается в следующем: из базы данных выбираются символы, отношения высоты к ширине которых находятся в некоторой окрестности. У текущего символа измеряются длины отрезков — расстояния от сторон и углов описывающего прямоугольника до изображения символа. Далее длины отрезков текущего символа сравниваются с отрезками из базы данных. Расстояние между символами измеряется как евклидово расстояние между полученными отрезками.

Точность описанных выше методов оказалась недостаточной, поэтому была применена модификация метода «сравнения контуров», приведенного в [Belongie et al., 2002]. Модификация метода заключается в следующем. На контуре символа случайным образом выбирается 100 точек. Для каждой точки строится круговая гистограмма, состоящая из 60 областей. Подсчитывается количество точек, попавших в каждую из областей. Расстояние между множествами точек на двух изображениях ищется как хи-квадрат. С помощью "венгерского метода" подбирается такая комбинация попарной связи точек, которая минимизирует суммарное расстояние. Это расстояние и явля-

ется расстоянием между двумя изображениями. Проведенные эксперименты показали, что расстояние между одинаковыми символами колеблется от 200 до 500, между похожими от 450 до 900, между разными от 900. Примеры приведены на рис.1.

Кластеризация полученных изображений осуществлялась методом иерархического кластерного анализа, используя полученную таблицу расстояний между выделенными изображениями. Таким образом, были выделены 395 кластеров. Общее количество кластеров превышает число используемых стенографических символов.

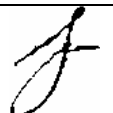
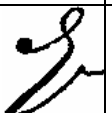
Одинаковые символы		Расстояние	Похожие символы		Расстояние	Разные символы		Расстояние
		405			547			1222
		297			662			1343

Рис. 1. Результаты сравнения контуров

При распознавании стенограмм появилась проблема выделения строк. Выделение строк является одной из ключевых проблем в расшифровке стенограмм, так как значение символа зависит от его места в строке. Так как стенограммы писались на нелинованных листах, с большим количеством исправлений и зачеркиваний, а так же с искривлением строки, в тексте сложно выделить явно выраженные строки.

Для выделения строк были испробованы методы, основанные на проекции изображения на вертикальную ось и на поиске символа, ближайшего к уже найденному символу в строке. Ни один из выше перечисленных методов не дает приемлемый результат на нашей коллекции документов. Поэтому для выделения строк мы использовали комбинированный алгоритм на основе поиска ближайшего символа в рамках строк, полученных сочетанием алгоритмов проекции черных пикселей символов и центров символов. Данный алгоритм был обучен на 100 строках, выделенных из пяти документов, расшифрованных вручную.

Авторами разработан прототип автоматизированной системы распознавания рукописных исторических документов в виде Web-приложения. Система имеет 3 рабочих области: область оригинального изображения стенограммы, область разобранной стенограммы и область с расшифрованным текстом стенограммы, а также всплывающее окно с возможными вариантами расшифровки символа.

При выделении фрагмента с символом на исходном изображении стенограммы, выделенный фрагмент бинаризуется, сегментируется и выделяется символ. Во второй области символ отображается в соответствующем месте в уже разобранной стенограмме. Далее, в третьей области отображается де-

шифрованный текущий символ, с возможными альтернативными вариантами дешифрования. Система, анализируя исходное изображение при вводе символов, производит автоматическое дешифрование схожих символов или групп символов.

Список литературы

Рогов и др., 2011 — Рогов А.А., Скабин А.В., Талбонен А.Н., Штеркель И.А. Некоторые особенности создания автоматизированной системы дешифровки исторических стенограмм // Интернет и современное общество: сборник научных статей. Материалы XIV Всероссийской объединенной конференции «Интернет и современное общество». Санкт-Петербург, 12–14 октября 2011 г. СПб., 2011. С. 132–138.

Ольхин, 1866 — Ольхин П. Руководство к русской стенографии. СПб., 1866.

Zhang and Suen, 1984 — T.Y.Zhang, C.Y.Suen. A fast parallel algorithm for thinning digital patterns // Commun. ACM. 1984. Vol. 27. № 3, Pp. 236–239.

Belongie et al., 2002 — S.Belongie, J.Malik, J.Puzicha. Shape matching and object recognition using shape contexts // IEEE Transactions on Pattern Analysis and Machine Intelligence. Apr 2002. Vol. 24. № 4. Pp. 509–522.