# АВТОМАТИЗАЦИЯ ДЕШИФРОВКИ ИСТОРИЧЕсКИХ СТЕНОГРАММ

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

Петрозаводский государственный университет

Петрозаводск

rogov@psu.karelia.ru

В настоящее время проводится множество исследований в направлении распознавания исторических документов. Особенно востребованной является возможность дешифровки стенографических документов. Исследование методов автоматизированной дешифровки стенограмм актуально по ряду причин. Главной причиной является отсутствие специалистов, способных дешифровать стенограммы прошлого века. В течение XIX и начала XX века стенография активно развивалась. Появилось множество различных стенографических систем, которые существенно отличаются от используемых в настоящее время. Таким образом, для дешифровки архивов необходимо обучать специалистов, что требует больших временных и материальных затрат.

Цель нашего исследования заключается в создании универсальной программной системы для автоматизированного распознавания исторических рукописных текстов, включая исторические стенограммы XIX и начала XX века. Она призвана решить задачу описания и дешифровки исторических стенограмм, а также ввести в научных оборот новые документы. Данное исследование поддержано грантом РГНФ ¹ 11-01-12026в (рук. Рогов А. А.).

Отличительные свойства разрабатываемой системы: учет особенностей исторической орфографии XIX и начала XX века, учет индивидуальных знаков разных стенографистов, возможность критического анализа, использование словаря для подсказки при дешифровке текста и т.д. [1]. Информационная система будет находиться в открытом доступе и предлагаться к использованию работникам архивов, научным сотрудникам, исследователям текстологам. Отлаживание системы происходит на стенограммах А. Г. Снит–киной. Выбор именно этих стенограмм обусловлен тем, что часть их была расшифрована.

Сложности при распознавании текста создают искривления строк, перепады яркости, просвечивания текста с обратной стороны и другие дефекты оригинала и изображения. Распознавание рукописного текста создает дополнительные сложности в отличие от распознавания печатного текста [2].

В рамках исследования решается задача создания достаточно универсальной программной системы для автоматизированного распознавания исторических стенограмм, для которых автоматическое распознавание оказывается пока невозможным. Предлагаемая система автоматизированной дешифровки исторических стенограмм с возможностью интеллектуальной поддержки принятия решений при наборе позволит существенно ускорить процесс перевода рукописного текста в текстовый файл и повысит точность его дешифровки.

**Бинаризация исторических рукописных документов**

При расшифровке стенограммы изображение переводится в бинарный вид с помощью порогового метода. Вследствие старости документов возникает проблема с бинаризацией их изображений. Пороговый метод по цветовым компонентам (RBG) оказался неприемлемым для данной задачи, так как пиксели фона и символов имеют схожие значения цветовых компонент. Как видно на гистограммах (рис. 1, а), отсутствие



Рис. 1. Гистограммы цветовых схем RGB и HSB

двух явно выраженных пиков не позволяет выбрать пороговое значение для бинаризации. Такие же результаты получаются (рис. 1, б), если использовать разложение по цветовой схеме HSB (оттенок, насыщенность, яркость). Производя бинаризацию только по пороговому значению яркости, мы смогли получить четкие символы с малым количеством шума. При этом была решена задача нахождения оптимального порогового значения яркости, при которой символы получаются наиболее четкими, с наименьшим количеством шума. Экспериментальным путем было выявлено, что наилучший результат достигается, если процент черных пикселей после бинаризации приближается к 13% от общего числа пикселей.

**Модуль создания оригинальной графики символов**

Система разбита на несколько модулей. Один из них – это модуль создания оригинальной графики символов. На рис. 2 представлен интерфейс модуля создания оригинальной графики символов.

Основное окно модуля представляет собой две области. В левой области находится оригинальное изображение (оригинальная стенограмма), на которой пользователь выделяет необходимый символ (рис. 2). Место выделения отображается на второй области. В этой области находится обработанная стенограмма, т.е. все полученные ранее символы, находящиеся в местах, соответствующих символам в оригинальном изображении.

После выделения символа пользователь должен выбрать пункт меню «Определить символ» или нажать на сочетание «горячих клавиш». Далее система производит бинаризацию выделенного фрагмента и его сегментацию. Система предлагает пользователю выбрать сегмент или сегменты, соответствующие оригинальному символу. Если было выбрано несколько сегментов, то система производит связывание разорванных «кусков». В случае когда результат устраивает пользователя, символ записывается в базу знаний и располагается в правой области соответственно месту (координатам) на оригинальном изображении. Если результат не соответствует требованию пользователя, то предусмотрена возможность редактирования полученного символа при помощи упрощенного графического редактора.

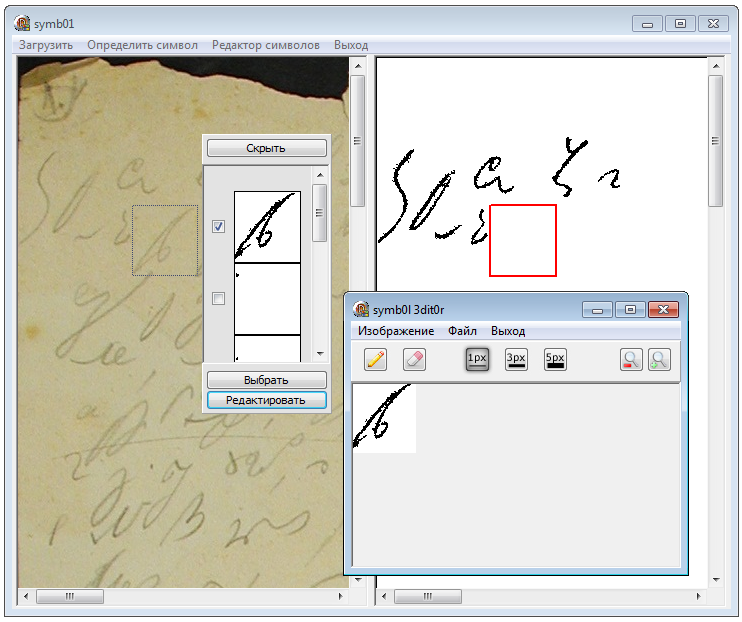


Рис. 2. Интерфейс модуля создания оригинальной графики символов

**Методы поиска и распознавания графем**

Существует множество способов поиска графем на исходном изображении. При этом они обеспечивают различные показатели точности и времени выполнения. Нами было принято решение об использовании комбинации методов распознавания. На первом этапе исходное изображение сравнивается с эталоном графемы с помощью вычисления длин контрольных отрезков. На выделенном пользователем изображении символа проводятся 8 отрезков, как показано на рис. 3.

  
Рис. 3. Расчет контрольных отрезков

Полученные значения сравниваются со значениями контрольных отрезков графем в БД. Данный метод не требует проведения сложных вычислений и поэтому позволяет быстро выделить множество наиболее вероятных соответствий графем выделенному символу. Следующий этап представляет собой проверку полученного множества соответствий более точным алгоритмом распознавания – попиксельным сравнением. Использование комбинации методов позволяет существенно повысить скорость распознавания с минимальными потерями точности.

**Прототип автоматизированной системы распознавания рукописных   
исторических документов**

На рис. 4 представлен интерфейс прототипа автоматизированной системы распознавания рукописных исторических документов. Как видно на этом рисунке, система имеет четыре области, в том числе область с оригинальным изображением, область виртуальной клавиатуры с оригинальной графикой символов, которые были получены, используя модуль описанный ранее.

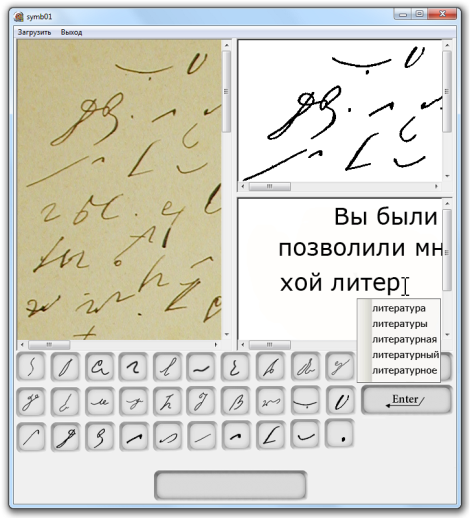


Рис. 4. Интерфейс прототипа автоматизированной системы распознавания   
рукописных исторических документов

При нажатии на кнопку виртуальной клавиатуры, изображение символа располагается в области 3, в том же месте, где он и находится на оригинальной стенограмме. В четвертой области отображаются дешифрованные символы. Так же система в процессе набора слова анализирует его составные части и предлагает пользователю близкие по значению расшифровки из базы знаний. Система, анализируя исходное изображение при вводе символов, производит автоматическое дешифрование схожих символов или групп символов.

В дальнейшем планируется реализация системы распознавания рукописных исторических документов в виде веб-сервиса для организации распределенной, удаленной работы со стенограммами.

**Библиографический список**

1. Рогов А. А., Талбонен А. Н., Варфоломеев А. Г. Автоматизированная система распознавания рукописных исторических документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: труды XII Всерос. науч. конф. RCDL’2010 (Казань, Россия, 13–17 октября 2010 г.). Казань: Казан. ун-т РАН, 2010. С. 469–475.
2. Горский Н., Анисимов В., Горская Л. Распознавание рукописного текста: от теории к практике. СПб.: Политехника, 1997.