

**Тезисы доклада**

Начало формы

1. **НАЗВАНИЕ ДОКЛАДА:**

Методы поиска схожих изображений стенографических символов.

Search methods of similar images of symbols from shorthand reports.

1. **АВТОРЫ:**

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

A. A. Rogov, A. V. Skabin, I. A. Shterkel

1. **ОРГАНИЗАЦИЯ (полное наименование, без аббревиатур):**

Петрозаводский государственный университет

Petrozavodsk State University

1. **ГОРОД:**

Петрозаводск

Petrozavodsk

1. **ТЕЛЕФОН:** (+78142) 71-96-45
2. **ФАКС:**
3. **E-MAIL**: shterkel\_ivan@petrsu.ru
4. **АННОТАЦИЯ:**

В докладе рассматривается распознавание рукописных исторических документов, представленных в цифровом виде. Приведено множество методов распознавания рукописного текста и описание наиболее эффективных при работе со стенографическими документами.

This report considers the recognition of historical shorthand documents which are presented in a digital form. A set of methods of recognizing handwritten texts is mentioned. Those that are more efficient in the work with shorthand reports are described.

1. **КЛЮЧЕВЫЕ СЛОВА:**

Распознавание, классификация, рукописный текст, стенограмма

Recognition, classification, handwritten text, shorthand report

1. **ТЕКСТ ТЕЗИСОВ ДОКЛАДА:**

Проблема понимания текстов - один из важнейших вопросов истории. В связи с этим одним из бурно развивающихся научных направлений стало распознавание рукописных исторических документов, в частности стенограмм. Несмотря на то, что проблемами автоматического распознавания текста занимаются уже несколько десятилетий, они все равно актуальны.

Распознавание изображений включает следующие этапы:

* предварительная обработка изображения – перевод изображения в бинарный вид и удаление шумов;
* сегментация – выделение объектов текста: строк, слов, символов;
* анализ обработанных и сегментированных изображений символов – определение схожести изображений символов на основе значений выбранных признаков.

Далее подробно рассматривается анализ изображений символов. Предполагается определение схожести изображений символов, в рамках ранее полученной коллекции. Изображения символов коллекции были предварительно обработаны и сегментированы. В ходе обработки 29 листов стенограмм А.Г. Сниткиной было получено более 5000 изображений. Из полученных изображений были выделены пять контрольных выборок с различными символами. Тестирование методов сравнения символов проводилось на выделенных контрольных выборках.

Для решения задачи сравнения символов были проанализированы следующие методы сравнения символов (более подробно см. [2]):

* логическое сравнение с эталоном;
* сравнение со скелетом эталона;
* метод сравнения расстояний;
* метод моментов;
* сравнение проекций;
* сравнение форм.

В ходе испытаний выяснилось, что точность большинства методов, приведенных выше, снижается при росте числа возможных символов. При стенографическом написании документов используется большой набор символов, поэтому большинство приведенных выше методов, показали крайне плохой результат.

Методы, применимые для определения схожести стенографических символов:

1. Сравнение форм.

Случайным образом выбираются N точек изображения символа. Изображение делится на K областей (назовем их корзинками), как показано на рис.1(с)[4]. Для каждой точки происходит подсчет попаданий остальных точек в корзинки. В результате получаем N гистограмм.

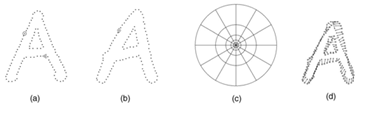


рис. Cравнение изображений по форме

Сравнение гистограмм различных точек производится по критерию X2.

|  |  |
| --- | --- |
|  |  |

Значения, полученные при сравнении гистограмм, образуют матрицу. Полученная матрица используется в качестве исходных данных для задачи назначений. В результате работы Венгерского метода формируются связи между точками (см. рис.1 (d)). На основании связей рассчитывается расстояние R (значение схожести).

|  |  |
| --- | --- |
|  | *π(i) – i-й вариант назначений* |

2. Сравнение проекций

Метод заключается в сравнении проекций точек изображений символов на оси X и Y(см. рис. 2)[5].

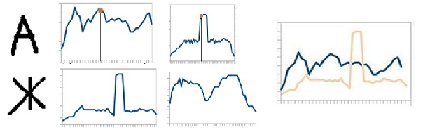


рис. Сравнение проекций

3. Сравнение длин отрезков

Сначала происходит построение контрольных отрезков на изображениях. Затем расстояние между символами вычисляется как эвклидово расстояние между парами полученных отрезков.

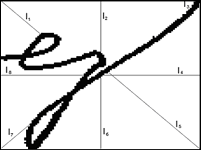


рис. Краевые расстояния

Каждый из описанных методов дает приемлемые результаты при поиске схожих изображений на контрольных выборках. Но поиск схожих изображений в полной коллекции приводит к снижению точности. Поэтому было принято решение о том, что необходимо учитывать результаты данных методов в совокупности. Таким образом, строится линейная комбинация полученных значений схожести, в рамках которой каждый метод имеет свой вес (голос).

Работа выполняется при финансовой поддержке Программы стратегического развития ПетрГУ в рамках реализации комплекса мероприятий по развитию научно-исследовательской деятельности.

Литература.

1. Скабин А.В., Рогов А.А. Бинаризация и выделение символов исторической стенограммы // Ученые записки Петрозаводского государственного университета. Серия «Естественные и технические науки». –2013. –№ 4 (133)
2. Рогов А.А. Скабин А.В. Штеркель И.А. Автоматизация дешифровки исторических стенограмм // Информационная среда ВУЗа XXI века: материалы V международной научно-практической конференции (26-30 сентября 2011 г.) – Петрозаводск: Петрозаводский государтвенный университет, 2011. - С. 164-168
3. Ольхин П. Руководство к русской стенографии. - СПб.: Типография доктора М. Хана, 1866
4. Belongie, S.; Malik, J.; Puzicha, J.; , "Shape matching and object recognition using shape contexts," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.24, no.4, pp.509-522, Apr 2002
5. Дробков А. В., Семенов А. Б. Обзор и анализ распознавателей рукопечатных символов // Математические методы распознавания образов (11-17 сентября 2011) – Тверь: Тверской Государственный Университет, 2011
6. Горский Н., Анисимов В., Горская Л. Распознавание рукописного текста: от теории к практике. – СПб.: Политехника, 1997