Март, № 2 **Физико-математические науки** **2014**

УДК 004.93

АЛЕКСАНДР АЛЕКСАНДРОВИЧ РОГОВ
доктор технических наук, профессор, заведующий кафедрой теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
rogov@petrsu.ru

ИВАН АЛЕКСАНДРОВИЧ ШТЕРКЕЛЬ
аспирант кафедры теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
shterkel_ivan@petrsu.ru

СРАВНЕНИЕ МЕТОДОВ РАСПОЗНАВАНИЯ ОБЪЕКТОВ НА СТЕНОГРАФИЧЕСКИХ ИЗОБРАЖЕНИЯХ*

Рассматриваются методы распознавания бинарных изображений с целью их сравнения и выделения наилучшего. Тестирование проводилось на коллекции стенографических документов XIX века. Изображения документов получены с помощью фотоаппарата. Качество низкое. Изображения коллекции были сегментированы и прошли бинаризацию. В итоге было получено 5823 символа. Для тестирования методов распознавания из полученной коллекции была выделена контрольная выборка из 234 изображений. В работе приведено описание 3 методов сравнения изображений, их достоинства и недостатки при распознавании рукописных стенографических символов. Полученные результаты позволили определить наилучший метод – метод сравнения форм. С помощью данного метода были получены меры сходства символов для всей коллекции, которые позволили провести кластеризацию коллекции стенографических символов. В итоге коллекция была разделена на 423 класса.

Ключевые слова: рукописные символы, сравнение методов распознавания, стенографические документы, Ф. М. Достоевский, А. Г. Сниткина

ВВЕДЕНИЕ

В современном научном сообществе большое внимание уделяется проблеме автоматизированного распознавания текста. Несмотря на то, что исследования по проблеме оптического распознавания текстов (optical character recognition, OCR) ведутся уже несколько десятилетий, данное направление остается крайне актуальным. В общем виде OCR – это задача перевода изображений текста в текстовые данные, используемые для представления символов в компьютере. Такая постановка задачи является трудноразрешимой, поэтому разработка OCR систем обычно нацелена на распознавание объектов в конкретно заданной области, например, распознавание банковских чеков или почтовых адресов [1]. В данной статье рассматриваются методы распознавания рукописного текста применительно к стенографическим документам, которые относятся к офлайн-методам. Сравнение методов распознавания было проведено на исторических стенограммах [4] А. Г. Сниткиной текстов Ф. М. Достоевского.

ОБЪЕКТ ИССЛЕДОВАНИЯ

При проведении офлайн-распознавания сначала решается задача предварительной обработки изображения, которая состоит из двух частей: 1) бинаризация; 2)сегментация.

В нашем случае бинаризация проводилась пороговым методом с оптимизацией его выбора [5]. Затем изображение было очищено от шумов (рис. 1). Сегментация проводилась автоматизированно. В результате обработки была сформирована коллекция, состоящая из 5823 символов.

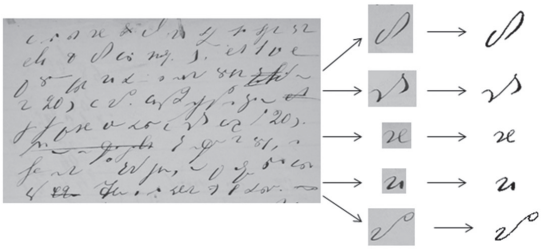

Рис. 1. Фрагмент стенограммы и обработанные символы

Главная особенность коллекции символов заключается в том, что из-за плохого качества изображений стенограмм после обработки появились разрывы символов. Для проведения исследований из общего количества символов было отобрано 234 и создана контрольная выборка.

Признаковые классификаторы основаны на значениях векторов признаков, с помощью которых определяется схожесть изображений.

Сравнение длин отрезков. В качестве меры сходства данный метод использует суммарную разницу между длинами отрезков, построенных по заранее определенным правилам. При сравнении стенограмм соблюдались следующие правила построения: отрезки строятся из угловых то-

чек, а также из середин отрезков, расположенных на границах изображения, в его центр, до первого касания с точками изображения [4]. Разница между длинами соответствующих отрезков складывается и служит мерой близости. Чем она меньше, тем более схожи изображения. Данный метод отличается высокой скоростью распознавания. Существенным недостатком является снижение точности при увеличении объема алфавита (набора оригинальных символов). Заметим, что данный метод чувствителен к искажениям и разрывам.

Сравнение проекций. Для сравниваемых изображений строятся графики проекций точек изображения на горизонтальную и вертикальную оси. Расстояние между изображениями определяется как суммарная разница между графиками вертикальной и горизонтальной осей [2]. Пример сравнения символов приведен на рис. 2.

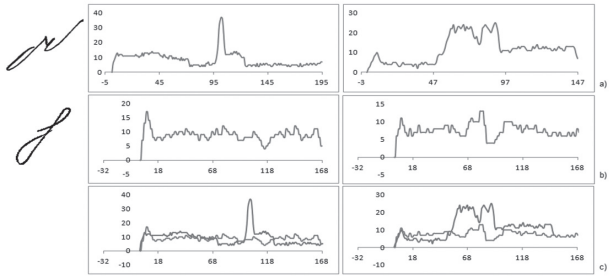

Рис. 2. Примеры проекций точек двух изображений: a) проекции 1-го изображения; b) проекции 2-го изображения; c) сравнение проекций

Вычисления, производимые при сравнении изображений методом проекций, достаточно просты и обеспечивают высокую скорость распознавания. Данный метод, так же как и метод проекций, чувствителен к искажениям и становится менее эффективен при увеличении размерности алфавита.

Структурные классификаторы переводят изображение символа в его топологическое представление, отражающее информацию о взаимном расположении структурных элементов символа.

Метод сравнения форм. Метод основан на определении положения точек изображения относительно друг друга. Случайным образом выбираются N точек изображения символа. Для каждой точки пространство вокруг нее делится на зоны (корзинки), как показано на рис. 3c. Оставшиеся точки, число которых N – 1, распределяются по корзинкам. Примем число корзинок равным K. В результате для каждого изображения получаем массив значений размерности N · K.

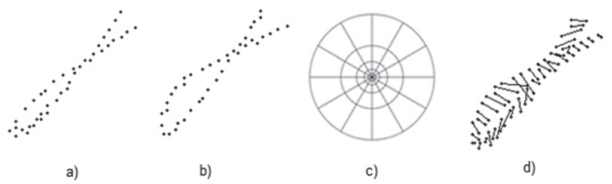

Рис. 3. Метод сравнения форм: a) и b) точки сравниваемых изображений, c) границы корзинок, d) связанные точки

Для того чтобы найти меру сходства изображений, нам необходимо найти суммарное смещение N точек одного изображения относительно N точек другого. При этом точки изображений сопоставляются с помощью решения задачи назначений [6]. Стоимость соединения точек мы определяем на основании распределения точек по корзинкам с помощью критерия X^2.

$$C_{ij} = C(p_i, q_j) = \frac{1}{2}\sum_{k=1}^{K}\frac{[h_i(k) - h_j(k)]^2}{h_i(k) - h_j(k)},$$

где $h_i(k)$ – число точек в k-й корзинке для i-й точки, $i = 1 \ldots N$, $k = 1 \ldots K$.

В качестве исходных данных задачи назначений мы получаем матрицу C со значениями C_{ij}, где $i, j = 1 \ldots N$. Задача назначений решалась Венгерским методом.

$$H(\pi) = \sum_i C(p_i, q_{\pi(i)}),$$

$$\sum_i C(p_i, q_{\pi(i)}) \to \min.$$

В результате получаем сопоставление выбранных N точек двух изображений. За меру сходства принимается суммарное Евклидово расстояние между этими точками. Данный метод устойчив к разрывам, но требует проведения большого числа вычислений.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

На основе анализа приведенных методов для решения задачи распознавания стенографических символов было решено, что наиболее пригодным является метод сравнения форм. Качество распознавания методов определялось на контрольной коллекции. Пример результатов работы методов приведен в табл. 1. Изображения приведены в порядке возрастания расстояния по сравнению с эталоном. Метод сравнения форм – 1, сравнения длин отрезков – 2, сравнения проекций – 3.

Таблица 1

Поиск символов методами сравнения расстояний, сравнения проекций и сравнения форм

1	эталон	1072	1255	1438	1486	1700	2020	2034	2035	2094	2097
2		12	56	60	66	69	70	77	83	88	88
3		829	837	870	879	905	917	934	938	946	952

Используя контрольную коллекцию, были получены оценки точности, полноты и F-меры, которые приведены в табл. 2. Наилучшие результаты показал метод сравнения форм.

Применение метода сравнения форм для решения задачи кластеризации. Как уже отме-

Таблица 2

Оценки эффективности методов

	Точность	Полнота	F-мера
1	54 %	93 %	0,684
2	48 %	83 %	0,606
3	47 %	86 %	0,615

чалось, метод сравнения форм обладает большой трудоемкостью и поэтому его использование для решения задачи кластеризации вызывает определенные трудности. Однако применение современных информационных технологий, в частности облачных, позволяет справиться с этой трудностью. Для проверки данного утверждения был проведен эксперимент. Коллекция, предназначенная для кластеризации, состояла из 5823 символов. Сложность каждого сравнения двух символов соответствует O (n^4). Для получения мер сходства между всеми символами необходимо было провести $\sum_{i=1}^{5822} i = 16950753$ операций сравнения символов. В рамках облака была организована вычислительная система, сравнения были распределены по 16 потокам. Расчет мер сходства длился 34 часа. В результате коллекция была разбита на 423 класса.

ЗАКЛЮЧЕНИЕ

Полученные результаты распознавания стенографических символов подтвердили сложность данной задачи. Наилучшие результаты распознавания были получены при использовании метода сравнения форм. Его точность составила 54 %. В дальнейшем авторы предполагают использовать метод комитетов, основанный на нескольких методах распознавания.

* Работа выполнена при поддержке Программы стратегического развития ПетрГУ в рамках реализации комплекса мероприятий по развитию научно-исследовательской деятельности на 2012–2016 гг.

СПИСОК ЛИТЕРАТУРЫ

1. Горский Н., Анисимов В., Горская Л. Распознавание рукописного текста: от теории к практике. СПб.: Политехника, 1997. 126 с.
2. 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. С. 350–353.
3. Местецкий Л. М. Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры. М.: ФИЗМАТЛИТ, 2009. 288 с.
4. Рогов А. А., Скабин А. В., Штеркель И. А. Методы поиска схожих изображений стенографических символов // Информационная среда вуза XXI века: Материалы VII Междунар. научно-практ. конф. Петрозаводск, 2013. С. 170–173.
5. Скабин А. В., Рогов А. А. Бинаризация и выделение символов исторической стенограммы // Ученые записки Петрозаводского государственного университета. Сер. «Естественные и технические науки». 2013. № 4 (133). С. 110–114.
6. Belongie, S., Malik, J., Puzicha, J. Shape matching and object recognition using shape contexts // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. Vol. 24. № 4. P. 509–522.

Rogov A. A., Petrozavodsk State University (Petrozavodsk, Russian Federation)
Shterkel' I. A., Petrozavodsk State University (Petrozavodsk, Russian Federation)

COMPARISON OF OBJECT RECOGNITION METHODS ON SHORTHAND REPORTS

The paper presents a set of recognition methods with the aim of comparing and selecting the best one. The research is based on a collection of shorthand reports of the XIXth century. The quality of images is low. Collected images were segmented and transferred into a binary form. Altogether, 5823 symbols were received. A control sample consisting of 234 images was selected from the obtained collection in order to test recognition methods. Three methods of image comparison are considered. Advantages and disadvantages of these methods aimed at recognition of shorthand symbols are presented. Based on the results obtained from methods' analysis, "Shape matching" was proposed as the best recognition method. The similarity measures were computed for all symbols. Subsequent clusterization of the studied collection was performed with the use of computed measures. As a result, the whole collection was divided into 423 classes.

Key words: optical character recognition, comparison of recognition methods, handwritten text, shorthand report, F. M. Dostoevsky, A. G. Snitkina

REFERENCES

1. Gorskiy N., Anisimov V., Gorskaya L. *Raspoznavanie rukopisnogo teksta: ot teorii k praktike* [Recognition of handwritten text: from theory to practice]. St. Petersburg, Politekhnika Publ., 1997. 126 p.
2. Drobkov A. V., Semyenov A. B. Review and analysis of hand printed symbols' recognizers [Obzor i analiz raspoznavateley rukopechatnykh simvolov]. *Matematicheskie metody raspoznavaniya obrazov (11–17 sentyabrya 2011)* [Mathematical methods of pattern recognition (11–17 september 2011)]. Tver': Tver state unversity Publ., 2011. P. 350–353.
3. Mestetskiy L. M. *Nepreryvnaya morfologiya binarnykh izobrazheniy: shapes, skeletons, circulars* [Continuous morphology of binary images: shapes, skeletons, circulars]. Moskow, Fizmatlit Publ., 2009. 288 p.
4. Rogov A. A., Skabin A. V., Shterkel I. A. Searching techniques of similar images of shorthand symbols [Metody poiska skhozhikh izobrazheniy stenograficheskikh simvolov]. *Materialy VII Mezhdunarodnoy nauchno-prakticheskoy konferentsii "Informatsionnaya sreda vuza XXI veka"* [Proc. 7th Int. Scientific Conference "Institution informational environment of higher education of XXI century"]. Petrozavodsk, 2013. P. 170–173.
5. Skabin A. V., Rogov A. A. Binarization and isolation of historical manuscripts' synbols [Binarizatsiya I vydelenie simvolov istoricheskoy stenogrammy]. *Uchenye zapiski Petrozavodskogo gosudarstvennogo universiteta. Ser. "Estestvennye i tekhnicheskie nauki"* [Proceedings of Petrozavodsk State University. Natural & Engineering Sciences]. 2013. № 4 (133). P. 110–114.
6. Belongie, S., Malik, J., Puzicha, J. Shape matching and object recognition using shape contexts // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002. Vol. 24. № 4. P. 509–522.

Поступила в редакцию 03.03.2014