**Сентябрь, № 6** **Физико-математические науки** **2015**

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**
программист РЦНИТ, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
*shterkel_ivan@petrsu.ru*

# ВЛИЯНИЕ ВЫБОРА РАСПРЕДЕЛЕНИЯ ТОЧЕК ИЗОБРАЖЕНИЯ ГРАФЕМ НА РАСПОЗНАВАНИЕ СИМВОЛОВ МЕТОДОМ СРАВНЕНИЯ ФОРМ*

Рассматривается распознавание изображений рукописных документов методом сравнения форм с целью нахождения оптимального алгоритма распределения точек изображения символа. В работе приведено описание трех методов выбора точек изображения графем и их влияния на качество распознавания. Каждый из методов протестирован на контрольной выборке из коллекции изображений стенографических документов размерностью 300 символов. Изображения представлены в бинарном виде. Полученные результаты позволили определить наиболее подходящий алгоритм выбора точек для распознавания методом сравнения форм. Выбор точек с использованием структурных аспектов их расположения в символе показал наилучшие результаты.

Ключевые слова: распознавание рукописных символов, распределение точек, метод сравнения форм, стенографические документы

## ВВЕДЕНИЕ

Одним из актуальных направлений исследований в области компьютерного зрения является распознавание рукописных документов. Существует множество решенных практических задач, например распознавание записей на почтовых отправлениях, банковских чеках, распознавание протоколов о происшествиях [1], [2]. Одной из востребованных задач является распознавание стенографических документов. Ее решение вызывает определенные сложности. Во-первых, в России в архивах находится большое число стенографических документов XIX века, которые не были расшифрованы, а специалистов, владеющих знаниями о существовавших стенографических системах, нет. Наличие расшифрованных документов позволяет с помощью системы распознавания автоматизировать процесс дальнейшей расшифровки стенограмм. Во-вторых, часто качество документов недостаточно высокое. Встречаются повреждение бумаги, выцветание записей, исправления текста, заваливание строк [3]. Все это создает проблемы при обработке изображений и влияет на точность распознавания.

## РАСПОЗНАВАНИЕ СИМВОЛОВ

В рамках работы были проведены следующие этапы [4]:

- сегментация символов стенограмм;
- бинаризация символов;
- очистка от шумов и мусора.

Пример обработки символов показан на рис. 1. В результате обработки была получена коллекция более чем из 5000 изображений стенографических символов. Для исследования влияния распределения выбранных точек изображения на качество работы методом сравнения форм была подготовлена контрольная выборка из 300 символов.

В качестве основного метода сравнения стенографических символов нами был выбран метод сравнения форм [5], [6]. Метод основан на определении положения точек изображения относительно друг друга. Количество точек сравниваемых изображений должно быть одинаковым. Обозначим множество точек за N. Выбор точек изображения осуществляется по заданным правилам из множества D. В зависимости от выбора множества точек результат распознавания может измениться, так как оно определяет геометрическую структуру символа.

После выбора для каждой точки пространство вокруг нее делится на зоны (назовем их корзинками), как показано на рис. 2c. Оставшиеся точки, число которых N-1, распределяются по корзинкам. Примем число корзинок равным K.

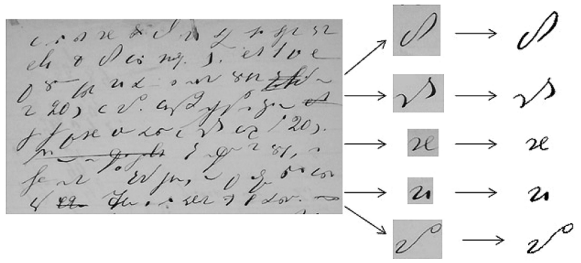

Рис. 1. Фрагмент стенограммы и обработанные символы

В результате для каждого изображения мы получаем массив значений размерности N*K.

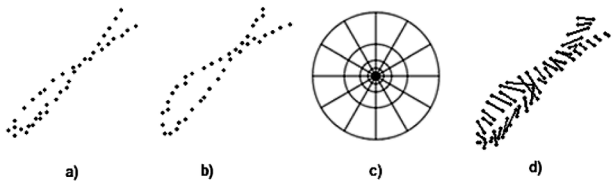


Рис. 2. Метод сравнения форм: a) и b) – точки сравниваемых изображений; c) границы корзинок; d) связанные точки

За меру сходства изображений примем суммарное смещение N точек одного изображения относительно N точек другого. Точки изображений сопоставляются с помощью решения задачи назначений (рис. 2d). Стоимость соединения точек мы определяем на основании распределения точек по корзинкам с помощью критерия $X^2$.

$$C_{i,j} = C\left(p_i, q_j\right) = \frac{1}{2} \bullet \sum_{k=1}^{K} \frac{\left[h_i(k) - h_j(k)\right]^2}{h_i(k) + h_j(k)},$$

hi(k) – число точек в k-й корзинке для i-й точки, где i = 1..N; k = 1..K.

В качестве исходных данных задачи назначений мы получаем матрицу C со значениями cij, где i,j = 1..N. Задача назначений решается венгерским методом.

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

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

В результате мы получаем сопоставление выбранных N точек двух изображений. За меру сходства принимается суммарное Евклидово расстояние между этими точками.

Рассмотрим различные варианты выбора точек.

**Выбор точек с использованием равномерной случайной величины**. Используем генератор случайных величин с дискретным равномерным распределением. На каждом шаге мы получаем номер точки и извлекаем точку из множества D. В итоге на n шаге получаем множество выбранных точек N. Данный алгоритм не требует большого числа вычислений, но качество распределения точек низкое.

**Выбор точек с использованием правил для обеспечения визуально равномерного распределения точек на плоскости.** Используем генератор случайных величин с дискретным равномерным распределением и правило позиционирования точек. Первый этап заключается в определении максимального расстояния между любыми двумя точками изображения. Максимальное расстояние, деленное на 10, будет контрольным расстоянием. Следующий этап предполагает последовательный выбор случайных точек. На каждом шаге выбирается точка, которая проходит проверку. Расстояние от данной точки до любой ранее выбранной точки должно превышать контрольное расстояние. Если точка проходит проверку, то она извлекается из множества D во множество N. Если оставшиеся в D точки не проходят проверку позиционирования, а необходимое число точек во множестве N не достигнуто, то уменьшается контрольное расстояние. В результате полученное множество N содержит максимально удаленные друг от друга точки. Данный алгоритм требует большее число вычислений, чем первый, что связано с проверкой правил и многократным обходом множества D. При этом полученные результаты равномерно распределены на плоскости изображения символа.

**Выбор точек с использованием структурных аспектов их расположения в символе.** Первый этап. Выбор точек осуществляется на основе их схожести по расположению в символе. Для каждой точки изображения мы строим множество корзинок и подсчитываем вхождения остальных точек в них. Обозначим число корзинок для каждой точки K. В результате мы получаем D множеств K. Следующий этап заключается в вычислении расстояния $X^2$ между всеми точками изображения. Полученные расстояния сортируются в порядке убывания. Обозначим множество расстояний X. На третьем этапе осуществляется выбор точек. Первая пара точек с самым большим расстоянием является исходной и имеет наиболее различные структурные показатели. Последующие пары выбираются из упорядоченного списка расстояний в порядке убывания. Важным фактором является проверка схожести структурных признаков не только самой пары, но и соотношения точек пары с ранее выбранными точками. Пусть ij – текущая пара для выбора. Тогда рассматриваем выбор точки i. Расстояние $R_{ij}$ между точками i и j берем как эталон на данном шаге. Далее определяем расстояние от точки i до каждой точки из множества N.

$$R_i = \min_{p=1..N} R_{ip}.$$

Если $R_i$ меньше, чем эталонное Rij, то точка не выбирается. Проверка точки j происходит аналогично. Выбор пары точек и их проверка происходит до наполнения множества N необходимым числом точек. Данный метод требует большого числа вычислений. При этом он учитывает структурные особенности написания символа, например толщину линий. Распределение точек получается визуально равномерным на плоскости.

## СРАВНЕНИЕ РЕЗУЛЬТАТОВ РАСПОЗНАВАНИЯ В ЗАВИСИМОСТИ ОТ ВЫБОРА РАСПРЕДЕЛЕНИЯ ТОЧЕК

На рис. 3 приведены изображения распределения точек для каждого из перечисленных методов.

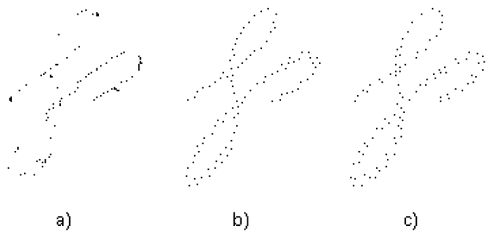


Рис. 3. Пример распределения точек

Результаты работы метода сравнения форм с различными распределениями точек на контрольной выборке приведены в таблице. Первая строка соответствует выбору точек с использованием равномерной случайной величины. Вторая – выбору точек с использованием правил для обеспечения визуально равномерного распределения точек на плоскости. Последняя строка содержит результаты работы метода сравнения форм с выбором точек с использованием структурных аспектов их расположения в символе.

Результаты работы метода сравнения форм

| | Точность | Полнота | F-мера |
|---|---|---|---|
| 1 | 41 % | 89 % | 0,561 |
| 2 | 54 % | 93 % | 0,684 |
| 3 | 54 % | 95 % | 0,688 |

Наилучший результат показал выбор точек с использованием структурных аспектов их расположения в символе. Данный метод имеет большую вычислительную сложность на этапе подготовки, но современные технологии, такие как распределенные вычисления и облачные технологии, позволяют достичь приемлемого времени обработки символов.

* Работа выполняется при финансовой 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. 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. Рогов А. А., Штеркель И. А. Сравнение методов распознавания объектов на стенографических изображениях // Ученые записки Петрозаводского государственного университета. Сер. «Естественные и технические науки». 2014. № 2 (139). С. 118–120.
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)

## THE INFLUENCE OF GRAPHEME IMAGE DOTS' DISTRIBUTION ON SYMBOL RECOGNITION BY THE SHAPE MATCHING METHOD

The paper is concerned with the shape and image recognition method of the shorthand reports' images. The method is studied with the aim of optimal algorithm determination for the symbol image dots' distribution. This work contains a description of three grapheme image dots selection methods and results of their influence on the effectiveness of recognition. The research is based on the collection of shorthand reports of the XIX$^{th}$ century. All collected images were segmented and transferred into a binary form. Every method was tested by the control image set. The size of the control image set contains 300 symbols. The obtained results assisted in determination of the most suitable dot selection algorithm for the shape and image recognition method. This method is a selection method based on structural features.

Key words: optical character recognition, dots' distribution, comparison of selection methods, shorthand reports

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., Semenov A. B. Review and analysis of hand printed symbols' recognizers [Obzor i analiz raspoznavateley rukopechatnykh simvolov]. *Matematicheskie metody raspoznavaniya obrazov* [Mathematical methods of pattern recognition]. Tver', Tver state university Publ., 2011. P. 350–353.
3. Mestetskiy L. M. *Nepreryvnaya morfologiya binarnykh izobrazheniy: figury, skelety, tsirkulyary* [Continuous morphology of binary images: shapes, skeletons, circulars]. Moscow, 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]. *Informatsionnaya sreda vuza XXI veka: Materialy VII Mezhdunarodnoy nauchno-prakticheskoy konferentsii* [Proc. 7th Int. Scientific Conference "Institution informational environment of higher education of XXI century"]. Petrozavodsk, 2013. P. 170–173.
5. Rogov A. A., Shterkel' I. A. Comparison of object recognition methods on shorthand reports [Sravnenie metodov raspoznavaniya ob''ektov na stenograficheskikh izobrazheniyakh]. *Uchenye zapiski Petrozavodskogo gosudarstvennogo universiteta. Ser. "Estestvennye i tekhnicheskie nauki"* [Proceedings of Petrozavodsk State University. Natural & Engineering Sciences]. 2014. № 2 (139). P. 118–120.
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 28.08.2015*