

АРТЕМ ВИКТОРОВИЧ СКАБИН

аспирант кафедры теории вероятностей и анализа данных математического факультета, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)  
artb00g@gmail.com

АЛЕКСАНДР АЛЕКСАНДРОВИЧ РОГОВ

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

## БИНАРИЗАЦИЯ И ВЫДЕЛЕНИЕ СИМВОЛОВ ИСТОРИЧЕСКОЙ СТЕНОГРАММЫ\*

Рассматривается проблема бинаризации исторических рукописных документов с целью выделения графики символов. От точности и правильности бинаризации текста зависит правильность результата дешифровки исторического документа. В качестве объекта взяты исторические стенограммы XIX века. Проводится анализ различных методов бинаризации (методы Отсу, Бернсена, Эйквилля, Ниблэка, разнообразные пороговые методы). Проведенное исследование выявило, что предлагаемый модифицированный пороговый метод показал лучший результат на основании F-меры. Данный метод применен в программном комплексе для получения оригинальной графики символов. При помощи разработанного программного комплекса были разобраны 29 листов стенограмм. На них было выделено более 6800 график символов.

Ключевые слова: методы бинаризации, исторические рукописные документы, стенограммы

### ВВЕДЕНИЕ

В настоящее время из-за невозможности дешифровки исторических документов современными исследователями в архивах России хранится большой объем нерасшифрованных стенографических документов. Основная сложность дешифровки стенограмм заключается в том, что современная стенография существенно отличается от исторических систем стенографии XIX века и нет людей, обладающих знаниями о системах стенографической записи в XIX – начале XX века. Наиболее естественной выглядит автоматизация данного процесса с использованием современных компьютерных технологий. Для автоматизации расшифровки исторических рукописных документов необходимо решить следующие задачи:

- выделение символов (как правило, это бинаризация документа);
- выделение полного набора стенографических символов (задача кластеризации);
- разделение связанных стенографических символов;
- база знаний дешифровки символов;
- выделение строк, а также надстрочных и подстрочных символов;
- дешифровка стенограммы.

### БИНАРИЗАЦИЯ СТЕНОГРАММЫ

Бинаризация исторических рукописных документов осложнена их плохим состоянием (пожелтевшая бумага, написаны простым карандашом, на бумаге существуют разрывы, грязь, просвечи-

вание с обратной стороны листа и т.д.). Для выбора наиболее эффективного метода бинаризации исторических стенограмм Сниткиной (наброски к «Дневнику писателя» за 1881 год и «Из воспоминаний А. Г. Достоевской») было рассмотрено 7 различных методов бинаризации. На рис. 1а представлен фрагмент одной из рассмотренных стенограмм, на рис. 1б – его эталонная бинаризация.

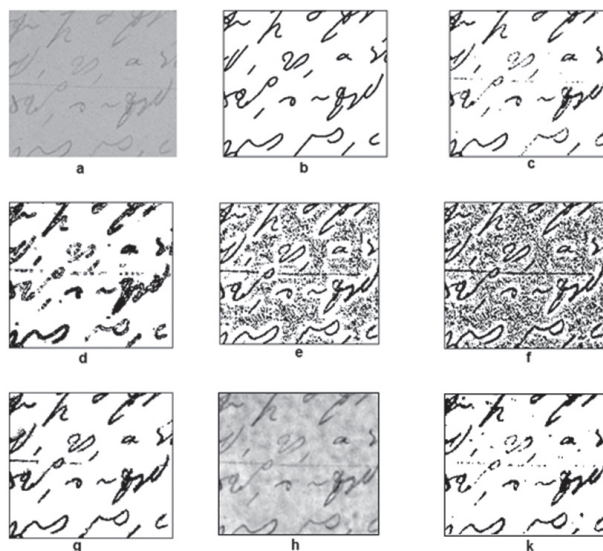


Рис. 1. а – фрагмент стенограммы, б – эталонная бинаризация, с – метод Отсу, d – метод Бернсена, е – метод Эйквилля, f – метод Ниблэка, g – модифицированный пороговый метод, h – Гауссовское размытие, k – результат модифицированного порогового метода на Гауссовском размытии

## МЕТОД ОТСУ

Метод Отсу [10] используется для пороговой бинаризации полутоновых изображений. При обработке изображений данным методом подразумевается наличие пикселей двух классов: фоновые и пиксели, относящиеся к текстовой информации изображения. Метод заключается в поиске оптимального порога бинаризации, который разделяет изображение на два вышеописанных класса таким образом, чтобы их внутриклассовая дисперсия была минимальной. Она определяется как взвешенная сумма дисперсий двух классов:

$$\sigma_w^2(t) = \omega_1(t)\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t), \quad (1)$$

где  $\omega_1, \omega_2$  – веса, вероятности двух классов, разделенных порогом  $t$ ,  $\sigma_1^2, \sigma_2^2$  – дисперсия этих классов. Результат работы метода Отсу представлен на рис. 1с.

## МЕТОД БЕРНСЕНА

Метод Бернсена [7] является локальным пороговым методом бинаризации, при котором пороги рассчитываются исходя из значений уровня яркости пикселей из квадратной окрестности пикселя  $p_{i,j}$ . Данный алгоритм заключается в следующем: для каждого пикселя  $p_{i,j}$  находится максимальное  $Y_{max}$  и минимальное  $Y_{min}$  значение уровня яркости. Яркость пикселя рассчитывается по формуле (2) [3]

$$Y = 0,299 \cdot R + 0,5876 \cdot G + 0,114 \cdot B, \quad (2)$$

где  $R$  – значение красной компоненты пикселя,  $G$  – значение зеленой компоненты пикселя,  $B$  – значение синей компоненты пикселя. Далее, если уровень контраста – разность между минимальным и максимальным значениями яркости – больше некоего порога, тогда данный пиксель относят к классу пикселей символов, иначе – к классу фона. Недостатком данного метода является то, что при обработке монотонных (по яркости) областей появляется большое количество помех и шумов. Результат работы метода Бернсена представлен на рис. 1д.

## МЕТОД ЭЙКВИЛЯ

Метод Эйквилья [8] является одним из самых производительных методов бинаризации. Он часто применяется для обработки четких и контрастных изображений. В данном методе используются два скользящих, обычно квадратных окна: малого  $S$  и большого  $W$ . Данные окна накладываются с шагом, равным ширине малого окна.

Для всех пикселей большего окна вычисляется оптимальный порог по методу Отсу, если разница математического ожидания уровней яркости в двух кластерах больше некоторого поро-

га, заданного пользователем  $|\mu_0 - \mu_1| \geq \varepsilon$ . Тогда пиксели малого окна бинаризируются в зависимости от найденного порога. Если же  $|\mu_0 - \mu_1| < \varepsilon$ , тогда все пиксели малого окна относятся к классу с ближайшим средним значением. Результат работы метода Эйквилья представлен на рис. 1е.

## МЕТОД НИБЛЭКА

Метод Ниблэка [9] является быстрым локальным адаптивным методом бинаризации. Зачастую данный метод используется без бинаризации контрастных изображений с плавными переходами яркости и малым количеством шума.

Пороговое значение для бинаризации пикселя  $p_{i,j}$  находится по следующей формуле:

$$t(i, j) = \mu(i, j) - k\sigma(i, j),$$

где  $\mu(i, j)$  – среднее,  $\sigma(i, j)$  – среднеквадратичное отклонение в окрестности пикселя  $p_{i,j}$ .

Размер окрестности для расчета порога бинаризации пикселя выбирается достаточно большой, чтобы снизить влияние локального шума, но не настолько, чтобы при обработке терялись локальные детали бинаризуемого изображения. Результат работы метода Ниблэка представлен на рис. 1ф.

## ПОРОГОВЫЕ МЕТОДЫ БИНАРИЗАЦИИ

Существует большое множество различных пороговых методов бинаризации изображений: бинаризация с нижним порогом, с верхним порогом, с двойным ограничением и т. д. Сложность использования данных методов заключается в правильном расчете порога бинаризации. Эти методы являются высокопроизводительными, но дающими большое количество шума и неточностей. Суть данных методов заключается в том, чтобы выбрать некий порог, такой, что:

$$p_{i,j} = \begin{cases} 0, & p_{i,j} \geq t \\ 1, & p_{i,j} < t \end{cases}.$$

## МОДИФИЦИРОВАННЫЙ ПОРОГОВЫЙ МЕТОД

Из-за состаренности изображения и того, что стенографические записи сделаны простым карандашом на пожелтевшей бумаге, пороговый метод по цветовым компонентам (RGB) оказался неприемлемым для данной задачи, так как пиксели фона и символов имеют схожие значения цветовых компонент. Как видно на гистограммах (рис. 2), отсутствие двух явно выраженных пиков не позволяет выбрать пороговое значение для бинаризации. Такие же результаты получаются (рис. 2), если использовать разложение по цветовой схеме HSB (оттенок, насыщенность, яркость). Производя бинаризацию только по пороговому значению яркости, удалось получить четкие символы с малым количеством шума [4].

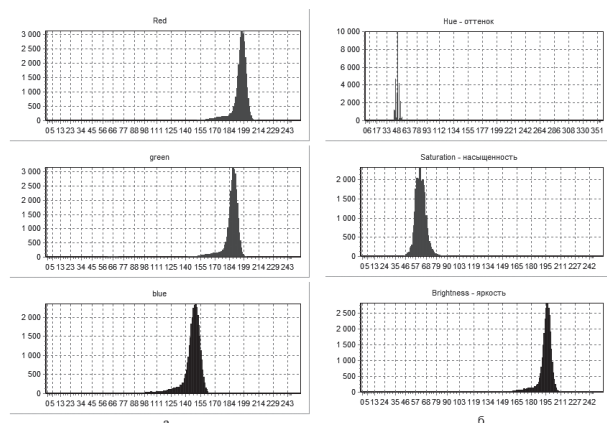


Рис. 2. Гистограммы цветовых схем RGB (а) и HSB (б)

Аналогичные результаты получены в [1] при распознавании банковских чеков.

Экспериментально было найдено пороговое значение яркости, при котором символы получаются наиболее четкими, с наименьшим количеством шума. Наилучший результат достигается, если доля черных пикселей после бинаризации приближается к 13 % от общего числа пикселей. Результат работы модифицированного порогового метода представлен на рис. 1г.

Для улучшения результатов бинаризации используют различные методы предобработки изображения: эвристический метод порогового отсечения без параметров, методы пространственной фильтрации, применяющие лапласиан, методы выделения границ, методы сглаживания [5]. В данной статье рассматривается размытие Гаусса, примененное к исходному фрагменту (рис. 1h). Результат работы модифицированного порогового метода на предобработанном изображении показан на рисунке 1к.

## РЕЗУЛЬТАТЫ СРАВНЕНИЯ МЕТОДОВ

Оценка результатов работы различных методов бинаризации исторических рукописных документов была проведена на основе критериев полноты, точности и F-меры [2]. Точность

Результаты сравнения методов бинаризации

| Метод                                                   | Точность бинаризации | Полнота бинаризации | F-мера   |
|---------------------------------------------------------|----------------------|---------------------|----------|
| Метод Отсу                                              | 0,8637               | 0,6684              | 0,753602 |
| Метод Бернсена                                          | 0,7362               | 0,7841              | 0,759395 |
| Метод Эйквилья                                          | 0,5111               | 0,7228              | 0,598789 |
| Метод Ниблэка                                           | 0,4398               | 0,7703              | 0,559917 |
| Пороговый модифицированный метод                        | 0,8165               | 0,7717              | 0,793468 |
| Пороговый модифицированный метод + Гауссовское размытие | 0,8284               | 0,7433              | 0,783612 |

считалась как количество правильно определенных черных пикселей относительно эталона к общему числу черных пикселей в эталоне. Полнота равна количеству правильно определенных черных пикселей относительно эталона к общему числу черных пикселей эталонного бинаризованного рукописного документа. Результаты оценки работы методов представлены в таблице.

Исходя из полученных значений F-меры пороговый модифицированный метод демонстрирует наилучшие показатели при бинаризации рукописных исторических документов.

## КОМПЬЮТЕРНАЯ СИСТЕМА БИНАРИЗАЦИИ И ВЫДЕЛЕНИЯ СИМВОЛОВ

Для распознавания исторических стенографических рукописных документов необходимо иметь весь набор символов, который использовал стенографист для записи. Для этого была создана автоматизированная система получения и разделения оригинальной графики символов. После запуска системы пользователь видит перед собой рабочую область (рис. 3), которая разделена на две части. Пользователь при помощи пункта меню «Файл» → «Загрузить» загружает оригинальную стенограмму для разбора. После загрузки она отображается в левой рабочей области, а в правой отображаются уже разобранные ранее символы. После этого пользователь в левой части выделяет прямоугольную область, содержащую разбираемый символ, оригинальную графику которого он хочет получить. Выделение дублируется в правой рабочей области. С помощью пунктов меню «Распознавание» → «Распознать символ» или же горячей клавишей начинается распознавание символа. Компьютерная программа с помощью модифицированного порогового метода бинаризирует выделенный фрагмент стенограммы, удаляет шумы и производит определение связанных фрагментов.

Пользователю предъявляется найденный символ (рис. 4). В случае, если символ разбивается на отдельные фрагменты, то пользователю предъявляются все фрагменты, а он выбирает те, которые, по его мнению, относятся к текущему символу, при этом система производит связывание разорванных фрагментов. Если же, по мнению пользователя, символ неполностью схож с первоначальным (произошло слияние, разрыв символа), то предусмотрена возможность его дополнительного редактирования (кнопка «Редактировать выбранные»). Пользователю представляется упрощенный графический редактор для символов.

После бинаризации символа система производит поиск в текущей базе данных похожих символов и предлагает пользователю выбрать из найденных или же добавить новый символ. Степень схожести определяется на основе комбинированного алгоритма. В настоящий момент

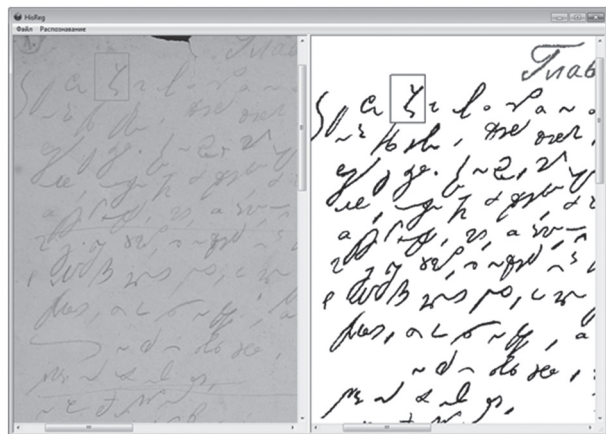


Рис. 3. Рабочая область системы получения оригинальной графики символов

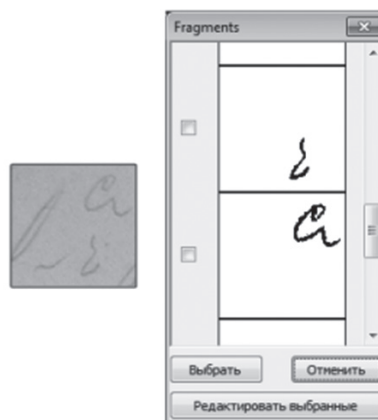


Рис. 4. Результаты обработки различных фрагментов стенограммы

мы работаем над его улучшением. При добавлении символа пользователь может указать зачеркнутые и нестенографические символы. Эта информация отображается в правой области цветом символов. Черный цвет означает стенографические символы, а красный – все остальные, включая зачеркнутые стенографистом символы.

Данная система получения оригинальной графики стенографических символов использовалась при обработке 29 листов стенограмм. На них было выделено более 6800 график символов, которые были разбиты с помощью модифицированного метода «сравнения контуров», описанного в [6], на более чем 400 групп (различные стенографические символы, буквы, цифры, слова). Проверить правильность разбиения можно только в процессе дешифровки, первые результаты которой обнадеживают.

## ЗАКЛЮЧЕНИЕ

В процессе анализа стенограмм было выявлено, что по цветовым компонентам значение пикселей текста и фона находятся в небольшой окрестности некоторого значения, что означает

низкую эффективность стандартных пороговых методов. Локальные пороговые методы, такие как Эйквилья и Ниблэка, при бинаризации дают хорошую полноту, но из-за большого количества шума, который не убирается при помощи удаления связанных компонент малой площади, точность их работы уменьшается. Наилучший результат работы показал модифицированный пороговый метод, при удалении шума он дает наибольшее значение F-меры. Повысить качество распознавания может предварительная обработка листов стенограмм [5]. В будущем мы планируем провести соответствующие эксперименты.

С компьютерной системой дешифровки стенограмм должен работать исследователь, поэтому одно из основных условий используемых методов в этой системе – быстроедействие. Необходимо создать комфортные условия работы с системой, пользователь не должен длительное время проводить в ожидании выполнения той или иной операции. Это условие накладывает существенные ограничения на используемые алгоритмы.

\* Работа поддержана грантом РГНФ № 11-01-12026в (руководитель – А. А. Рогов).

## СПИСОК ЛИТЕРАТУРЫ

1. Горский Н., Анисимов В., Горская Л. Распознавание рукописного текста: от теории к практике. СПб.: Политехника, 1997. 126 с.
2. Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск: Пер. с англ. М.: ООО «И. Д. Вильямс», 2011. 528 с.
3. Порев В. Н. Компьютерная графика. СПб.: БХВ-Петербург, 2002. 432 с.
4. Рогов А. А., Скабин А. В., Штеркель И. А. О дешифровке исторических рукописных документов // Информационные технологии и письменное наследие EIManuscript 2012: Материалы IV Междунар. науч. конф. Петрозаводск, 3–8 сентября 2012. Петрозаводск, 2012. С. 230–233.
5. Талбонен А. Н., Рогов А. А. Анализ машинописных подписей к фотографиям в цифровом историческом альбоме // Ученые записки Петрозаводского государственного университета. Сер. «Естественные и технические науки». 2012. № 2 (123). С. 109–113.
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.
7. Bernsen J. Dynamic thresholding of grey-level images // Proc. 8th International Conference on Pattern Recognition (ICPR8). Paris, 1986. P. 1251–1255.
8. Eikvil L., Taxt T., Moen K. A fast adaptive method for binarization of document images // Proc. 1st Int. Conf. Document Analysis and Recognition. Saint-Malo, 1991. P. 435–443.

9. Niblack W. An Introduction to Digital Image Processing. Englewood Cliffs, N. J.: Prentice Hall, 1986. P. 115–116.
10. Otsu N. A threshold selection method from gray-level histograms // IEEE Transactions on Systems, Man, and Cybernetics. 1979. Vol. SMC-9. P. 62–66.

**Skabin A. V.**, Petrozavodsk State University (Petrozavodsk, Russian Federation)

**Rogov A. A.**, Petrozavodsk State University (Petrozavodsk, Russian Federation)

## BINARIZATION AND ISOLATION OF HISTORICAL MANUSCRIPTS' SYMBOLS

The problem of historical manuscripts' binarization with the purpose of original symbol graphics' isolation is considered. Correctness of historical documents' decryption largely depends upon valid and proper binarization of the text.

Historical shorthand records of the XIX century are taken as objects of research. The analysis of various binarization methods (Otsu, Bernsen's methods, Eykvil, Niblek, and various threshold methods) was carried out. The conducted research proved that the proposed modified threshold method based on F-measure is more effective than other methods. The method is applied in the software complex aimed at getting original graphical representation of the symbols. The use of the elaborated software complex was instrumental in the reading of 29 manuscript sheets. More than 6,800 graphic symbols were isolated during decoding process.

**Key words:** binarization methods, historical handwritten documents, shorthand report

## REFERENCES

1. Gorskiy N., Anisimov V., Gorskaya L. *Raspoznavanie rukopisnogo teksta: ot teorii k praktike* [Handwriting recognition: from theory to practice]. St. Petersburg, Politekhnik Publ., 1997. 126 p.
2. Manning K. D., Raghavan P., Schuytse H. *Vvedenie v informatsionnyy poisk* [Introduction to Information Retrieval]. Moscow, OOO "I. D. Vil'yams" Publ., 2011. 428 p.
3. Porev V. N. *Komp'yuternaya grafika* [Computer graphics]. St. Petersburg, BHV-Peterburg-Publ., 2002. 432 p.
4. Rogov A. A., Skabin A. V., Shterke I. A. About deciphering Historical Manuscripts [O deshifrovke istoricheskikh rukopisnykh dokumentov] *Informatsionnye tekhnologii i pis'mennoe nasledie El'Manuscript 2012* [Informational technologies and written heritage]. Petrozavodsk, 2012. P. 230–233.
5. Talbonen A. N., Rogov A. A. Analysis of typewritten photo captions in digital historical records [Analiz mashinopisnykh podpisov k fotografiyam v tsifrovom istoricheskom al'bome]. *Uchenye zapiski Petrozavodskogo gosudarstvennogo universiteta. Ser. "Estestvennye i tekhnicheskie nauki"* [Proceedings of Petrozavodsk State University. Natural & Engineering Sciences]. 2012. № 2 (123). P. 109–113.
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.
7. Bernsen J. Dynamic thresholding of grey-level images // Proc. 8th International Conference on Pattern Recognition (ICPR8). Paris, 1986. P. 1251–1255.
8. Eikvil L., Taxt T., Moen K. A fast adaptive method for binarization of document images // Proc. 1<sup>st</sup> Int. Conf. Document Analysis and Recognition. Saint-Malo, 1991. P. 435–443.
9. Niblack W. An Introduction to Digital Image Processing. Englewood Cliffs, N. J.: Prentice Hall, 1986. P. 115–116.
10. Otsu N. A threshold selection method from gray-level histograms // IEEE Transactions on Systems, Man, and Cybernetics. 1979. Vol. SMC-9. P. 62–66.

*Поступила в редакцию 15.04.2013*