# О дешифровке рукописных исторических документов

© А. А. Рогов © А. В. Скабин © И. А. Штеркель
Петрозаводский Государственный Университет,
Петрозаводск
rogov@psu.karelia.ru artb00g@gmail.com shterkel_ivan@psu.karelia.ru

**Аннотация**

В статье описываются результаты исследования по созданию универсальной программной системы для автоматизированного распознавания исторических рукописных текстов, включая исторические стенограммы XIX и начала XX веков. Рассматривается проблема получения оригинальной графики символов исторических рукописных документов, поиск схожих график в базе знаний, определения строк текста, приводится математическая модель дешифровки стенограмм. Кроме этого описывается прототип автоматизированной системы распознания рукописных исторических документов.

*Данное исследование поддержано грантом РГНФ № 11-01-12026в.*

## 1 Введение

Проблема понимания текстов - один из важнейших вопросов истории. На сегодняшний день в архивах России накопился большой объем нерасшифрованных стенографических документов. Современные стенографисты и ученые испытывают существенные проблемы при дешифровке исторических стенограмм. Во-первых, в XIX и начале XX веков стенография в России находилась в процессе становления, поэтому существующие документы зашифрованы в разных системах. Современная стенография существенно отличается от исторических систем стенографии XIX века, и нет специалистов, обладающих знаниями о системах стенографической записи того времени. Еще одной трудностью дешифровки является то, что стенографист при шифровании мог использовать свои нестандартные символы (обозначения), так как часто расшифровкой занимался тоже он. Кроме того, некоторые символы стенографической записи имеют схожее написание, но в зависимости от некоторых физических параметров, например, таких как размеры, имеют разные значения.

**Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.**

Таким образом, для введения в научный оборот новых документов необходимо описание и дешифровка исторических стенограмм. Для решения этой задачи необходимо создать универсальную программную систему автоматизированного распознавания исторических рукописных текстов, включая исторические стенограммы XIX и начала XX веков. К отличительным свойствам разрабатываемой системы относятся: учет индивидуальных знаков разных стенографистов, возможность критического анализа, использование словаря для подсказки при дешифровке текста и т.д. [8]. Информационная система будет находиться в открытом доступе и предлагаться к использованию работниками архивов, научными сотрудниками, исследователями текстологам. Отлаживать систему было принято решение на стенограммах А.Г. Сниткиной, частично расшифрованных Ц. М. Пошемянской, и учебнике П. Ольхина [7].

Автоматизация дешифровки исторических стенограмм состоит из следующих шагов:

- бинаризация исторических рукописных документов;
- создание БД графики стенографических символов;
- кластеризация изображений стенографических символов;
- создание базы знаний стенографических символов;
- методы выделения строк в рукописных исторических документах;
- поиск символа в базе знаний;
- математическая модель распознавания символа.

## 2 Бинаризация исторических рукописных документов

Из-за состаренности рукописных материалов и того что стенографические записи сделаны простым карандашом на пожелтевшей бумаге возникает проблема с бинаризацией изображения. Пороговый метод по цветовым компонентам (RBG), оказался не приемлемый для данной задачи, так как пиксели фона и символов имеют схожие значения цветовых

компонент. Как видно на гистограммах (рис. 1) отсутствие двух явно выраженных пиков не позволяет выбрать пороговое значение для бинаризации. Такие же результаты получаются (рис. 1), если использовать разложение по цветовой схеме HSB (оттенок, насыщенность, яркость). Производя бинаризацию только по пороговому значению яркости, можно получить достаточно четкие символы, с малым количеством шума. Аналогичные результаты получены в [6] при распознавании банковских чеков.

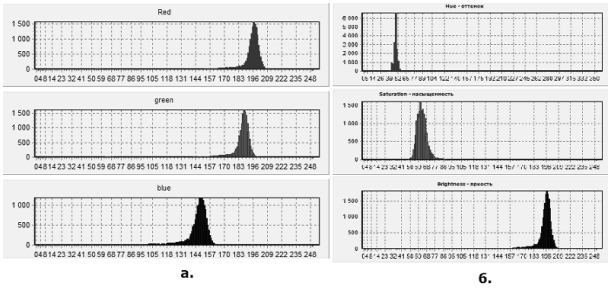



рис. 1 Гистограммы цветовых схем
RGB (а) и HBS (б)

Кроме того, был опробован метод бинаризации Оцу [3], но он так же как и предыдущие методы не давал хороших результатов при бинаризации, т.к. отсутствовали два явно выраженных пика. В нашем случае хорошее качество бинаризации дал метод с экспериментально найденным пороговым значением яркости 13% относительно общего числа пикселей.

## 3 База данных оригинальной графики стенографических символов

Для распознания стенограмм необходимы оригинальные символы, который использовал стенографист в своих записях, так как стенографист мог видоизменять и модернизировать символы стенографической системы, которую он использовал. Выделение символов на рукописных исторических документах имеет следующие особенности:

- оригинальное изображение удовлетворительного качества из-за следующего ряда причин:
  - рукописям порядка 150 лет;
  - запись производилась простым карандашом на пожелтевшей бумаге, которая имеет различные повреждения (перегибы, разрывы, просвечивание надписей с другой стороны листа);
  - наличие сторонних записей, пометок, которые не носят смысловой нагрузки, но пересекаются со стенографическими записями.
- при бинаризации происходили разрывы символов, т.к. некоторые пиксели символа имели схожий цвет с пикселями бумаги;
- при сегментации возникала необходимость разбиения символов, написанных слитно, на отдельные символы.

Было написано приложение позволяющее получать оригинальную графику символов из стенограмм. Как видно из рис. 2, на котором представлен интерфейс данного приложения, основное рабочее окно разделено на две части. В левой части находится оригинальное изображение стенограммы, из которой нужно получить графики символов. В правой части находятся уже полученные графики символов, причем они располагаются с полым соответствием расположения на оригинальном изображении.

Для создания оригинальной графики символа, пользователь должен выделить символ на оригинальном изображении и нажать на "горячую клавишу" или их комбинацию. Далее приложение производит бинаризацию и сегментацию выделенного фрагмента. Если при сегментации было получено несколько сегментов, то приложение предлагает пользователю выбрать, те, которые, по его мнению, соответствуют данному символу, и система производит связывание [4] "разорванных частей". Если пользователя не устраивает результат, то он может быть отредактирован во встроенном графическом редакторе. После этого система производит поиск в базе данных ранее полученных график символов, схожие, и предлагает пользователю выбрать, если нет схожих символ добавляется в базу данных.


рис. 2 Интерфейс модуля создания оригинальной графики символов.

Было обработано 29 листов стенографических записей и получено более 2500 изображений символов. После этого возникла задача их кластеризации.

## 4 Кластеризация изображений стенографических символов

Для решения этой задачи были проанализированы следующие методы сравнения символов:

- логическое сравнение с эталоном;
- сравнение со скелетом эталона;
- метод сравнения расстояний;
- метод моментов
- комбинированный метод;
- сравнения контуров.

Поиск с использование эталона заключатся в том, что изображение символа используется как шаблон и сравнивается с изображениями других символов. При сравнении изображений вычисляется расстояние Хэмминга.

$$\rho(\mathrm{x},\mathrm{y}) = \mathrm{D}(\mathrm{x},\mathrm{y}) \oplus G(\mathrm{x},\mathrm{y}), \mathrm{x} = 1..\mathrm{w}, \mathrm{y} = 1..h$$

Формально это означает сравнения значений каждого пикселя изображения первого символа со значением соответствующего пикселя второго символа. При этом значения принадлежат множеству {0,1}. Значения пикселей логически складываются, после чего происходит подсчет уровня совпадения, который равен числу нулей в полученной матрице, разделенному на общее число точек. Если полученный уровень превосходит заданный порог, то изображения считаются схожими.

При пороговых значениях уровня совпадения от 80% данный метод обладает высокой точностью поиска, но при этом полнота поиска крайне мала. Логическое сравнение с эталоном показывает плохие результаты для сложных символов, что происходит из-за того, что человек пишет символы по-разному. Меняются углы, размер и толщина символа. При сравнении скелетов символов исключается различие толщины символов, но усиливается различие в зависимости от других факторов написания символа. Скелетизация символов производилась алгоритмом Зонга Суня [5].

Метод сравнения расстояний является быстрым методом сравнения символов. Принцип данного метода заключается в построении отрезков по заранее заданному правилу и определение их длин. После вычисления длины заносятся в базу данных. В работе мы использовали метод краевых расстояний. Принцип работы метода краевых расстояний заключается в следующем, из базы знаний выбираются символы отношения высоты к ширине которых находится в некоторой окрестности. Это позволяет сократить множество символов, в котором осуществляется поиск. У текущего символа измеряются длины отрезков $\{l_1, l_2, \ldots, l_8\}$ (см. рис. 3). Далее длины отрезков текущего символа сравниваются с отрезками из базы данных. Расстояние между символами измеряется как эвклидово расстояние между полученными отрезками.

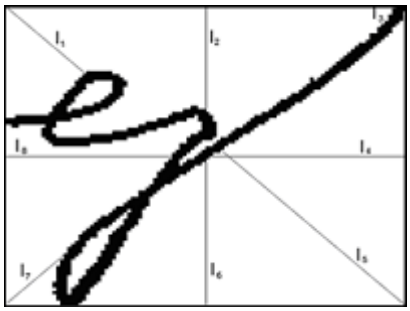


рис. 3 Метод краевых расстояний.

Проведенные испытания показали, что для большинства похожих символов исследуемой коллекции расстояние не превышает значения 100. Примеры приведены на рис. 4.

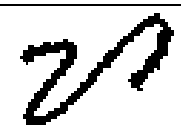
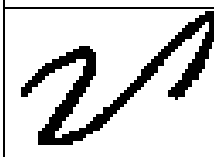
| Похожие символы | | расстояние |
|---|---|---|
|  |  | 23.6 |
|  |  | 57.3 |
| Разные символы | | расстояние |
|  |  | 121.8 |
|  |  | 128.2 |

рис. 4 Сравнение расстояний.

Метод моментов основан на вычислении расстояний между моментами изображений. Мы использовали инвариантные моменты Ху [2]. Рассмотрим процесс вычисления моментов Ху. Пусть f(x,y) функция интенсивности точек бинарного изображения, которая принимает значения «0» и «1» (где «0» – черный цвет, а «1» - белый). При сравнении изображений мы работаем с дискретными величинами, поэтому от непрерывного значения момента (1) мы переходим к дискретному случаю (2). Значение p – порядок x, q – порядок y.

$$M_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q f(x,y)\,dxdy, \quad p,q = 0,1,2,\ldots \quad (1)$$

$$M_{pq} = \sum_x \sum_y x^p y^q f(x,y) \quad (2)$$

Далее переходим к центральным моментам, которые инвариантны при сдвиге изображения.

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x, y) \quad (3)$$

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\left(1 + \frac{p+q}{2}\right)}} \quad (4)$$

Момент $\eta_{pq}$ инвариантен при масштабировании. В формуле (4) $\mu_{00}$ интерпретируется как полная масса изображения.

Ниже приведены моменты Ху, которые инвариантны при сдвиге, масштабировании и повороте.

$\varphi_1 = \eta_{20} + \eta_{02}$

$\varphi_2 = (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2$

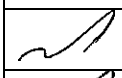
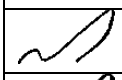
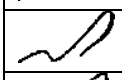
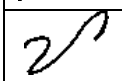
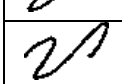
$\varphi_3 = (\eta_{30} - 3\eta_{12})^2 + (\eta_{21} - \mu_{03})^2$

$\varphi_4 = (\eta_{30} - 3\eta_{12})^2 + (\eta_{21} + \mu_{03})^2$

$\varphi_5 = (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2]$
$+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$

$\varphi_6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})$

$\varphi_7 = (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$
$- (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]$

| | Hu1 | Hu2 | Hu3 | Hu4 |
|---|---|---|---|---|
|  | 0,2147 | 0,0139 | 1,5925E-05 | 1,2413E-05 |
|  | 0,2094 | 0,0109 | 1,8696E-05 | 1,7896E-05 |
|  | 0,2176 | 0,0136 | 5,6078 | 4,3674E-05 |
|  | 0,2117 | 0,0053 | 3,0727E-05 | 9,3077E-06 |
|  | 0,1938 | 0,0034 | 5,6286E-06 | 4,7169E-06 |

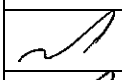
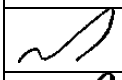
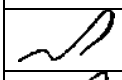
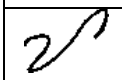
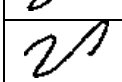
| | Hu5 | Hu6 | Hu7 |
|---|---|---|---|
|  | 1,7090E-10 | 8,5411E-07 | 3,5408E-11 |
|  | 3,1893E-10 | 1,2221E-06 | -7,3859E-11 |
|  | 2,1551E-09 | 3,6388E-06 | -1,6609E-10 |
|  | 1,5224E-10 | 1,6644E-07 | 3,9974E-11 |
|  | 2,4304E-11 | 7,9079E-08 | 1,5291E-13 |

рис. 1 Значения моментов

В ходе исследования были вычислены моменты Ху для коллекции стенографических символов. Как видно из рис. 5, разброс значений конкретных моментов для похожих символов может различаться на несколько порядков, в то время как значения различных символов могут быть значительно ближе. На основании полученных данных был сделан вывод, что моменты Ху не позволяют определить, похожи стенографические символы или нет.

Комбинированный метод представляет собой линейную комбинацию с весовыми коэффициентами, описанных выше методов. Точность описанных выше методов оказалась недостаточной и, поэтому, была применена модификация метода "сравнения контуров", приведенного в [1]. Примененная модификация метода заключается в следующем. На контуре символа случайным образом выбирается 100 точек. Для каждой точки строиться круговая гистограмма, состоящая из 60 областей. Подсчитывается количество, попавших точек из 100 выбранных, в каждую из областей. Расстояние между множествами точек на двух изображениях ищется как хи-квадрат. С помощью "венгерского метода" подбирается такая комбинация попарной связи точек, которая минимизирует суммарное расстояние. Это расстояние и является расстоянием между двумя изображениями. Проведенные эксперименты показали, что расстояние между одинаковыми символами колеблется от 200 до 500, между похожими от 450 до 900, между разными от 900. Примеры приведены на рис. 6.

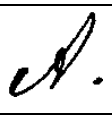
| Одинаковые символы | | расстояние |
|---|---|---|
|  |  | 405 |
|  |  | 297 |
| Похожие символы | | расстояние |
|  |  | 547 |
|  |  | 662 |
| Разные символы | | расстояние |
|  |  | 1222 |
|  |  | 1343 |

рис. 6 Метод сравнения контуров.

Кластеризация полученных изображений осуществлялась методом иерархического кластерного анализа, используя полученную таблицу расстояний между выделенными изображениями. Таким образом, были выделены 395 кластеров. Общее количество кластеров превышает число стенографических символов, так как текст содержит цифры и буквы (слова, записанные обычным образом).

## 5 Методы выделения строк в рукописных исторических документах

При распознании стенограмм появилась проблема выделения строк (линий). Выделение строк является одним из ключевых в расшифровки стенограмм, так как значение символа зависит от его места нахождения в строке. Символы могут быть: надстрочными, подстрочными, и обычные. Так как стенограммы писались на нелинованных листах, с большим количеством исправлений и зачеркиваний, а так же с искривлением строки, поэтому в тексте сложно выделить явно выраженные строки.

Для выделения строк были испробованы следующие методы:

- Проекция центров тяжести символов. Были спроецированы центры тяжестей символов, но не возможно было выделить ярко выраженные пики, т.к. строки имеют кривизну;
- Проекция черных пикселей символов. При проекции черных стали более выделены пики соответствующие строками (см. рис. 8 верхний). Но при искривлении строк, а так же при большом количество зачеркиваний и исправлений, некоторые строки не соответствуют пикам (см. рис. 8 нижний);
- Алгоритм поиска ближайшего символа в строке. Символ считается ближайшим, если символ находится от текущего в секторе 60 градусов с наименьшим расстоянием. Вес близости символа рассчитывается по формуле:

$$w = d \cdot k + l$$

где d – угол между прямой соединяющей центры символов и осью OX, k – коэффициент схожести , l – расстояние до символа (см. рис. 7). Из-за подстрочных и надстрочных символов некоторые символы не соответствовали строке или же происходит пропуск символов.

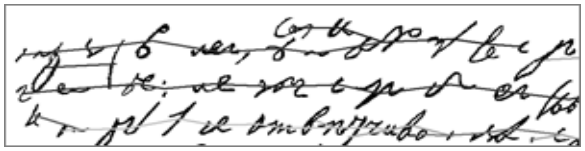


Рис. 7. Результат работы алгоритма поиска ближайшего символа

Не один из выше перечисленных методов, не дает правильный результат в случае с рукописными историческими документами. Для выделения строк используется комбинированный алгоритм на основе поиска ближайшего символа в рамках строк, полученных сочетанием алгоритмов проекции черных символов и центров символов. Данный алгоритм был обучен на порядке 100 строк, выделенных из пяти разобранных документов.

## 6 Математическая модель распознавания символа

Основной задачей в расшифровке документов – это сопоставление каждому символу его значение, но каждый символ может иметь несколько значений, либо же схожие по написанию символы могут иметь совершенно разные значения. Для данной задачи была составлена математическая модель, но на данном этапе исследования, пока не достаточно информации для проверки эффективности её работы.

Обозначим через $x_1,\ldots,x_n$ последовательность стенографических символов. К сожалению, очень часто стенографические символы определяются неоднозначно. Для символа $x_k$ обозначим через $x_1^k,\ldots,x_{l_k}^k$ множество его возможных распознаваний. Каждому распознанному символу $x_i^k$ на основании БЗ ставиться в соответствии его возможные трактовки $y_1^{ki},\ldots,y_{m_{ki}}^{ki}$. Тогда распознанный текст примет вид $y_{j_1}^{1i_1},\ldots,y_{j_n}^{ni_n}$.

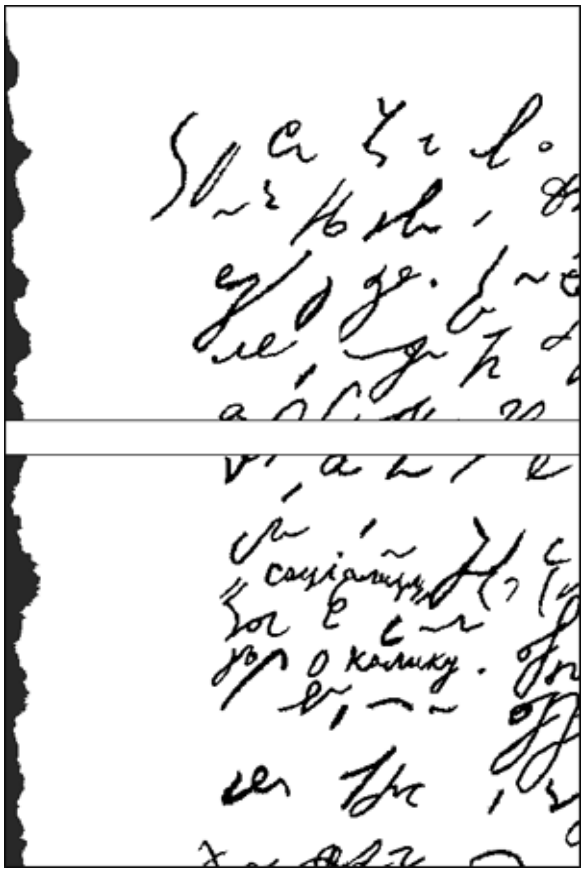


Рис. 8. Проекция черных пикселей символов

Ставится задача найти такой набор индексов, чтобы вероятность правильного распознавания была максимальной.

$P\left(y_{j_1^*}^{1i_{1}^*}\ldots y_{j_n^*}^{ni_{\text{n}}^*}\right)=\max\ P\left(y_{j_1}^{1i_1}\ldots y_{j_n}^{ni_{\text{n}}}\right)$, где максимум берется по всем $1\le i_1\le l_1,\ 1\le j_1\le m_{1i_1},\ldots,\ 1\le i_n\le l_n,$ $1\le j_n\le m_{ni_n}$. Оценим вероятность $P\left(y_{j_1}^{1i_1}\ldots y_{j_n}^{ni_{\text{n}}}\right)$. На основании формулы Байеса она равна

$$P\left(y_{j_1}^{1i_1}\ldots y_{j_n}^{ni_{\text{n}}}\right)=P\left(y_{j_1}^{1i_1}\right)\ldots P\left(y_{j_n}^{ni_{\text{n}}}\left|{}_{y_{j_1}^{1i_1}\ldots y_{j_{n-1}}^{(n-1)i_{\text{n-1}}}}\right.\right) \quad (1)$$

Оценка k-го (k>3) сомножителя в правой части формулы (1) имеет вид:

$$P\left(y_{j_k}^{ki_{\text{k}}}\left|{}_{y_{j_1}^{1i_1}\ldots y_{j_{k-1}}^{(k-1)i_{\text{k-1}}}}\right.\right)=aP\left(x_{i_k}^{k}\left|{}_{x_{i_{k-3}}^{k-3}\ldots x_{i_{k-1}}^{k-1}}\right.\right)+$$

$$(1-a)P\left(y_{j_k}^{ki_{\text{k}}}\left|{}_{y_{j_{k-3}}^{(k-3)i_{k-3}}\ldots y_{j_{k-1}}^{(k-1)i_{\text{k-1}}}}\right.\right) \quad (2)$$

Первое слагаемое в правой части формулы (2) характеризует точность распознавания стенографического символа. Оно вычисляется как взвешенная сумма:

$$P\left(x_{i_k}^{k}\left|{}_{x_{i_{k-3}}^{k-3}\ldots x_{i_{k-1}}^{k-1}}\right.\right)=b\frac{\max\limits_{1\le i\le l_k}R\left(x_i^k,x_k\right)-R\left(x_{i_k}^k,x_k\right)}{\max\limits_{1\le i\le l_k}R\left(x_i^k,x_k\right)}+$$

$$(1-b)\frac{N\left(x_{i_{k-3}}^{k-3}\ldots x_{i_k}^{k}\right)}{N\left(x_{i_{k-3}}^{k-3}\ldots x_{i_{k-1}}^{k-1}\right)+1}. \quad (3)$$

В первом слагаемом правой части формулы (3) $R\left(x_{i_k}^k,x_k\right)$ означает расстояние между символом и его возможным эталонным значением. Во втором слагаемом правой части формулы (3) $N\left(x_{i_{k-3}}^{k-3}\ldots x_{i_k}^{k}\right)$ означает частоту появления комбинации символов $x_{i_{k-3}}^{k-3}\ldots x_{i_k}^{k}$ в стенограммах.

Второе слагаемое в правой части формулы (2) характеризует вероятность появления данного фрагмента текста. Оно оценивается как

$$P\left(y_{j_k}^{ki_{\text{k}}}\left|{}_{y_{j_{k-3}}^{(k-3)i_{k-3}}\ldots y_{j_{k-1}}^{(k-1)i_{\text{k-1}}}}\right.\right)=$$

$$\frac{N\left(y_{j_{k-3}}^{(k-3)i_{k-3}}\ldots y_{j_k}^{ki_{\text{k}}}\right)}{N\left(y_{j_{k-3}}^{(k-3)i_{k-3}}\ldots y_{j_{k-1}}^{(k-1)i_{\text{k-1}}}\right)+1},$$

где $N\left(y_{j_{k-3}}^{(k-3)i_{k-3}}\ldots y_{j_k}^{ki_{\text{k}}}\right)$ частота появления фрагмента текста $y_{j_{k-3}}^{(k-3)i_{k-3}}\ldots y_{j_k}^{ki_{\text{k}}}$. Данная оценка производится на основании анализа текстов автора, в данном случае Ф.М. Достоевского. Оценка k-го сомножителя при $k\le 3$ производится аналогично. Коэффициенты $a,b$ настраиваются в зависимости от качества распознавания стенограммы.

## 7 Прототип Web-приложения системы автоматизированной системы распознания рукописных исторических документов

На рис. 9 представлен интерфейс прототипа автоматизированной системы распознавания рукописных исторических документов в виде Web-приложения. Система имеет 3 рабочих области. Область оригинального изображения стенограммы, область разобранной стенограммы, область с расшифрованным текстом стенограммы, а также всплывающее окно с возможными вариантами расшифровки символа.

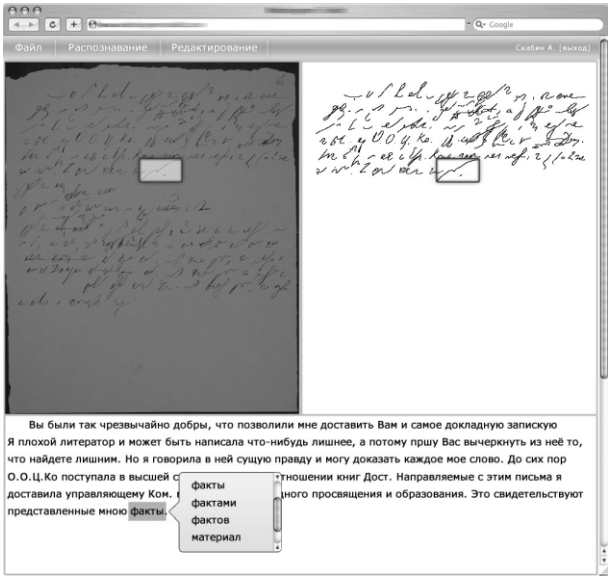



рис. 9 Интерфейс прототипа Web-приложения автоматизированной системы распознания рукописных исторических документов

При выделении символа на оригинальной стенограмме, выделенный фрагмент бинаризуется, сегментируется и выделяется символ. Во второй области отображается в соответствующем месте. Далее в третьей области отображается дешифрованный текущий символ, с возможными альтернативными вариантами дешифрования. Система, анализируя исходное изображение при вводе символов, производит автоматическое дешифрование схожих символов или групп символов.

Данное web-приложение находится на этапе прототипа, в котором реализованы основные этапы получения графики символа для его расшифровки.

## 8 Заключение

На данном этапе разработки автоматизированной системы распознания исторических рукописных

документов было реализовано и получено следующее:

- Был реализован и внедрен модуль получение оригинальной графики символов, использованных при написании стенограмм Сниткиной. Было обработано порядка 25 стенограмм и было получено более 3 тысяч график символов;
- Был разработан алгоритм для выделения строк в стенограммах, но данный алгоритм требует универсализации, т.к. корректность выделения строк зависит от подбора параметров, которые различны на различных стенограммах;
- Разработанная математическая модель дешифровки текста проходит апробирование, которое затруднено тем, что словарь значений стенографических символов в настоящий момент находится на этапе наполнения. Наполнение происходит исходя из правил описанных в учебнике [7], по которому обучалась Сниткина;
- Создан прототип системы в виде web-приложения. Данный прототип не является окончательным и самостоятельным программмным продуктом, он требует доработки, оптимизации, для удобной работы пользо-вателей с ним.

## Литература


[1] 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

[2] Hu M.K. Visual pattern recognition by moment invariants. / M.K. Hu // IRE Transactions on Information Theory 8 − 1962 – P. 179–187.

[3] N. Otsu (1979). «A threshold selection method from gray-level histograms». *IEEE Trans. Sys., Man., Cyber.* 9: 62-66

[4] P Nagabhushan, Basavaraj S Anami. A knowledge-based approach for recognition of handwriting Pitman shorthand language storkes. / P Nagabhushan, Basavaraj S Anami // *Sadhana. – 2002.* - Vol. 27, Part 5. -P. 685–698.

[5] Zhang, T.Y. A fast parallel algorithm for thinning digital patterns / T. Y. Zhang, C. Y. Suen // Commun. ACM. − 1984. − Vol. 27, №3. – P. 236-239.

[6] Горский Н., Анисимов В., Горская Л. Распознавание рукописного текста: от теории к практике. – СПб.: Политехника, 1997 г.

[7] Ольхин П. Руководство к русской стенографии. - СПб.: Типография доктора М. Хана, 1866 г

[8] Рогов А.А., Скабин А.В., Талбонен А.Н., Штеркель И.А. Некоторые особенности создания автоматизированной системы дешифровки исторических стенограмм //Интернет и современное общество: сборник научных статей. Материалы XIV Всероссийской объединенной конференции «Интернет и современное общество». Санкт-Петербург, 12-14 октября 2011 г. – СПб.: Из-во ООО «МультиПроЖектСистемСервис», 2011. – С. 132-138.


## About the the decoding of handwritten historical documents


Aleksandr Rogov, Artem Skabin, Ivan Shterkel



The article describes the process of creating a universal computerized recognition system of historical manuscripts, including historical shorthand records dating back to the 19th and early 20th centuries. We discuss the problem of getting the original graphical representation of symbols from historical manuscripts using a threshold digitization method. We search for a similar graphical representation of symbols in the database. Moreover we present a prototype of a computerized recognition system of historical manuscripts.