КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 004.93’12

РАСПОЗНАВАНИЕ РУКОПИСНЫХ ОТДЕЛЬНЫХ СИМВОЛОВ РУССКОГО АЛФАВИТА С ПРИМЕНЕНИЕМ МЕТОДА ВЫЯВЛЕНИЯ «ХАРАКТЕРИСТИК» СИМВОЛА

Н. Е. Тимофеева, А. С. Гераськин

Саратовский национальный исследовательский государственный университет им. Н. Г. Чернышевского

Поступила в редакцию 09.08.2016 г.

Аннотация. Представлен метод распознавания рукописных отдельных символов русского алфавита на основе выявления набора «характеристик» символа и последующего сравнения этого набора с наборами характеристик из базы приложения. Дан алгоритм расстановки маркеров для получения характеристик символа. Согласно программному эксперименту различимы все рукописные символы из алфавита русского языка.
Ключевые слова: распознавание, отдельные рукописные символы, русский алфавит, маркеры, характеристики символа.
Annotation. This article presents the methods for the separate handwriting Russian alphabets recognition. It is based on the detection of characteristics of alphabets set and on the comparison this set with set of characteristics from the application database then. Also this article presents the algorithm of the alignment markers for getting characteristics of alphabets. In accordance with the program experiment all handwriting Russian alphabets are recognizable.
Keywords: ecognition, separate handwriting alphabets, Russian alphabet, marker, characteristics of alphabets.

ВВЕДЕНИЕ

С развитием компьютерной техники и внедрением ее во все сферы повседневной жизни стало удобно хранить различные виды информации в электронной форме. Изображения, видео, тексты, таблицы – хранение всех этих видов информации в цифровом виде открывает широкие возможности для длительного хранения, быстрого поиска и обмена информацией. На текущий момент изображения и видеоинформация часто сразу записывается в цифровой форме с помощью специальных цифровых фото и видеокамер. С их же помощью можно переснять в цифровой формат старые материалы того же вида. По-другому обстоит дело с текстовой информацией. В этом виде хранится большая часть всех обрабатываемых и хранимых данных. Однако текст может быть представлен в разных видах. Хорошо, если это печатный текст, набранный на компьютере. Тогда он сразу может быть представлен в цифровой форме. Но что делать с распечатанным на бумаге текстом, или, еще хуже, с текстом, написанным от руки? Вот тут и встает проблема распознания текста. На текущий момент существует множество приложений, которые позволяют с высокой точностью распознавать печатный текст. Хуже обстоит дело с документами, написанными от руки. Эффективного средства распознания таких текстов до сих пор нет. Над этой проблемой работают специалисты ведущих мировых компаний, в том числе и Microsoft Inc.

Цель данной работы состоит в разработке надежного метода распознавания рукописных отдельных символов русского алфавита, с применением метода выявления «характеристик» символа.

МЕТОДЫ РАСПОЗНАВАНИЯ РУКОПИСНЫХ МЕТОДОВ

Выделяют три основных класса методов применяемые для решения задачи распознавания символов: шаблонный, признаковый и структурный [1, 2].

Шаблонные методы. Этот класс осуществляет преобразование исходного изображения к растровому, и затем сравнивает его по точечно со всеми имеющимися в базе шаблонами. Данный тип методов обладает достаточно высокой устойчивостью к дефектам изображения, а также высокой скоростью обработки входных данных. Главным недостатком таких методов является то, что они способны эффективно распознавать символы только тех шрифтов, шаблоны которых есть в базе данных.

Структурные методы. Они представляют символ в виде графа, в котором, множество вершин составляют структурные единицы исходного изображения, а множество ребер – пространственные отношения между ними. Под структурными единицами в данном случае подразумеваются составляющие символ линии. Недостатками структурных методов является их высокая чувствительность к дефектам изображения, которые нарушают целостность структурных единиц. Векторизация, разбиение на структурные единицы и их анализ требует достаточно больших объемов памяти, а также большого количества процессорного времени.

Признаковые методы. Как понятно из названия признаковые методы анализируют не сами символы, а набор присущих им определенных признаков. Однако это же является и недостатком этих методов. Так как объект заменяется своим упрощенным представлением, то большая часть информации об изображении символа теряется. Как следствие – уменьшается вероятность однозначного определения символа. Повышается вероятность неопределенности или появления неверного результата [3].

МЕТОД ВЫДЕЛЕНИЯ ХАРАКТЕРИСТИК СИМВОЛА

Важной особенностью технологии распознавания рукописных символов, является невозможность использования в явном виде шаблонных и структурных методов распознавания, которые представляют собой наиболее мощный и точный механизм определения печатных букв. Из-за различий в написании одного и того же символа разными людьми, и даже одним человеком в различных условиях, сравнение символов простым наложением на образцы будет малоэффективно. Эффективность структурных методов снижается не столь сильно, но их реализация становится значительно более сложной задачей, чем в случае с печатными текстами. Меньше всего, при переходе к рукописным текстам, снижается эффективность признаковых методов распознавания, но данный метод изначально является менее точным, чем структурный или шаблонный. В связи с этим необходимо разработать метод, который будет в себе сочетать положительные стороны изложенных выше методов.

Суть разработанного метода заключается в выявлении набора «характеристик» символа и последующего сравнения этого набора с наборами характеристик из базы приложения. Получение и сравнение набора характеристик проходит в несколько этапов: расстановка маркеров, выявление характеристик, сравнение с эталонами.

РАССТАНОВКА МАРКЕРОВ

Расстановка маркеров будет осуществляться по разработанному алгоритму, представленному на рис. 1. Для начала изображение символа вписывается в прямоугольник, т. е. убирается пустое пространство с каждой из четырех сторон изображения. Затем оно подвергается процедуре перевода в черно-белую гамму и нормализации. Разбивается область

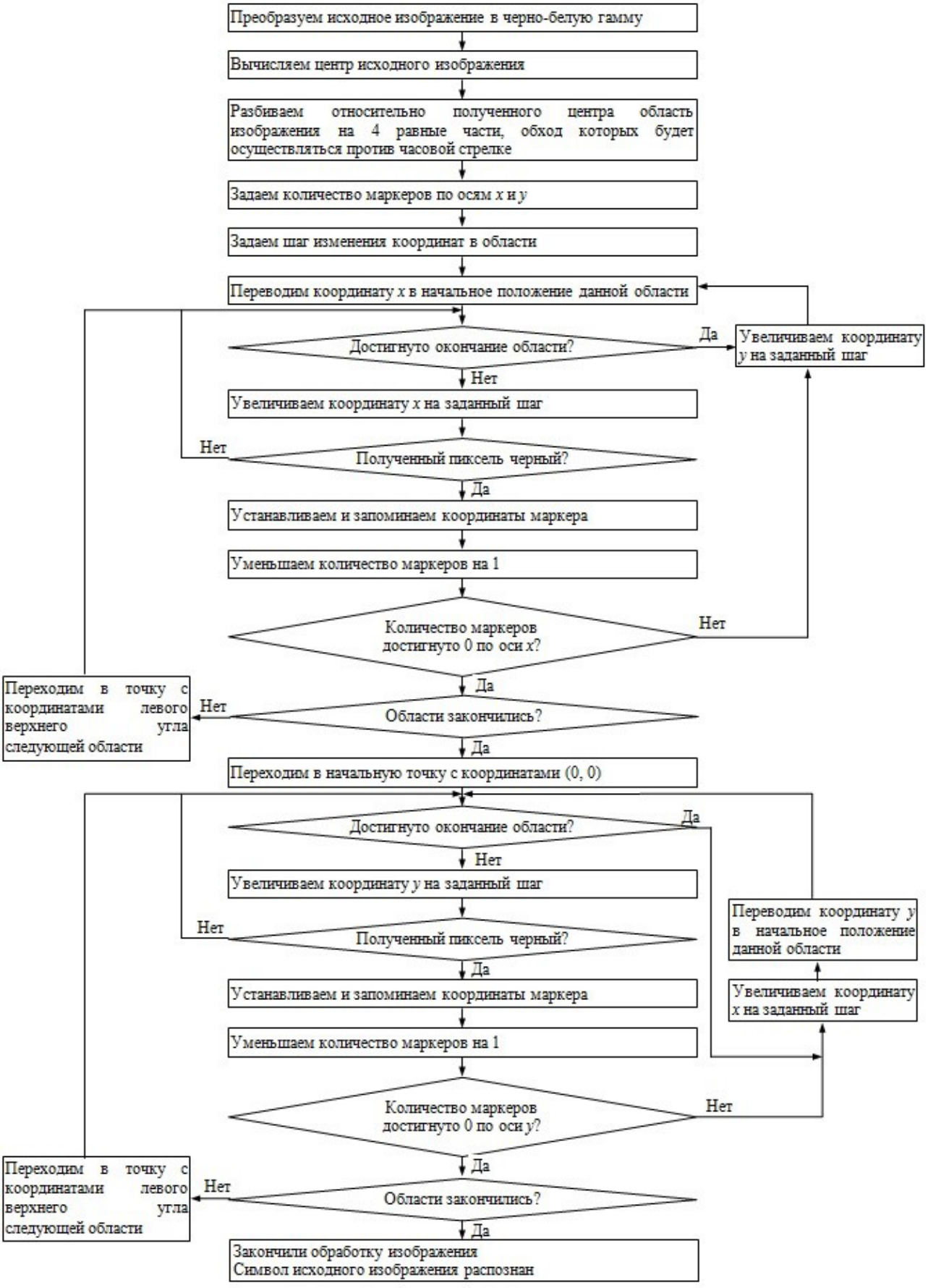

Рис. 1. Блок-схема алгоритма расстановки маркеров

на четыре части, обход которых происходит против часовой стрелки. В каждой получившейся области, расставляются заранее заданное количество маркеров через равные интервалы по осям X и Y. После расстановки, каждая точка направляется к противоположной стороне прямоугольника. Как только на ее пути попадается достаточно темный пиксель, точка останавливается на нем, и считаем, что данный пиксель принадлежит символу. По окончанию данной операции, внешняя поверхность символа становиться покрыта маркерами со всех сторон. Предварительное вписывание символа в прямоугольник обеспечивает отсутствие «пролетевших» маркеров вдоль каждой стороны, а нормализация цвета – большую точность в определении границ символа. Маркеры являют собой некоторое урезанное описание распознаваемого символа, но их точное положение будет сильно меняться при переходе от одного написания буквы к другому.

МЕТОДЫ СРАВНЕНИЯ ХАРАКТЕРИСТИК

Для выявления характеристик были применены следующие методы:

Сравнение маркеров по принципу «Всех из всех». Метод осуществляет сравнение координаты каждого маркера со всеми остальными, и запись относительного положения для каждой полученной упорядоченной пары маркеров. Относительность описания маркеров в характеристике, призвана освободить описание от привязки к координатам, размерам символа и снизить различие характеристик у различных начертаний одного и того же символа. К недостаткам первоначального метода можно отнести большой объем памяти для хранения характеристик символа. Метод при расстановке n маркеров, будем иметь O(n2) характеристик.

Построение «упорядоченных пар маркеров». Второй метод позволяет избавиться от хранения излишней информации. Суть этого метода заключается в том, что при подсчете характеристик, используются только пары маркеров вида (m[i], m[(i+1) mod n]). Таким образом, для каждого маркера записывается его положение, относительно следующего, в порядке обхода (последний сравнивался с первым). Этот подход снижает объем необходимой для хранения характеристик памяти до O(n).

Сравнение по группам. Этот метод был основан на методе Сравнение маркеров по принципу «Всех из всех» с некоторыми изменениями: все полученные маркеры делились на две группы – отправленные вдоль горизонтальной или вертикальной оси. Сравнение производилось только между маркерами из различных групп. Такой подход снижает объем памяти, необходимый для хранения характеристик.

РЕЗУЛЬТАТ ТЕСТИРОВАНИЯ МАРКЕРОВ

Для оценки эффективности каждого из методов, был использован следующий тест: исходно программе давался образец буквы (в тесте использовалась буква «к»), и несколько других символов, среди которых были несколько других начертаний буквы «к», а также другие символы, более и менее схожие по начертанию с «к». Тест рассчитывал схожесть представленных символов с образцом при различном количестве маркеров, запускаемых с каждой стороны. Результаты двух методов представлены в виде графиков на рис. 2–3.

Из приведенных графиков на рис. 2–3 можно сделать следующие выводы. Метод Сравнение маркеров по принципу «Всех из всех» демонстрирует стабильные результаты распознавания, практически для любого количества маркеров на сторону. Метод Построение упорядоченных пар показывает лучшие результаты на небольшом количестве маркеров. Лучший результат достигается при количестве маркеров равных от пяти до семи на сторону. Однако затем результативность этого метода быстро снижается, а начиная с определенного момента, результаты приобретают случайный характер. Это вызвано тем, что, начиная с некоторого момента, сравнение двух следующих друг за другом маркеров начинает характеризовать не форму распознаваемого символа, а неровность границы

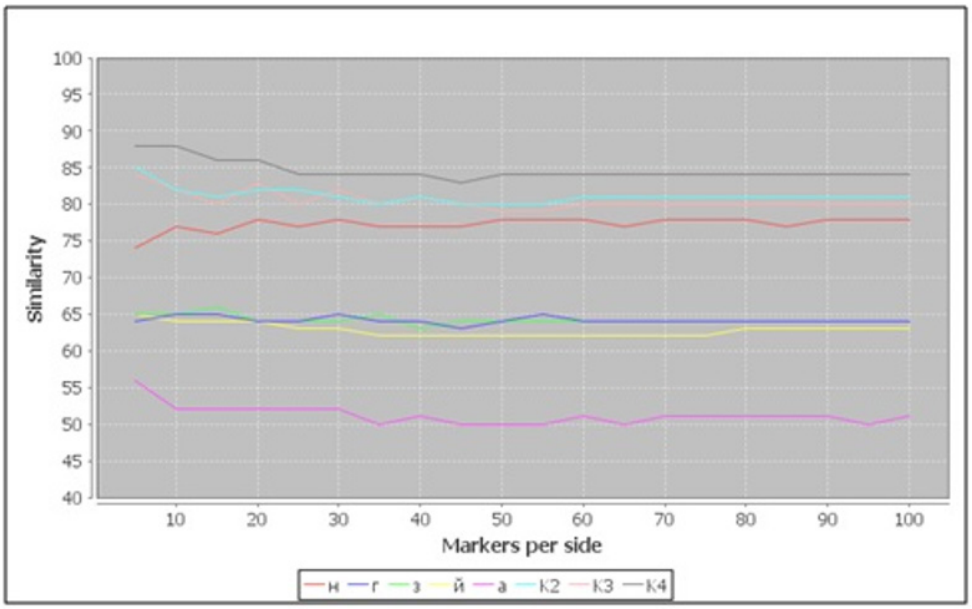

Рис. 2. График зависимости результатов сравнения от числа маркеров (метод Сравнение маркеров по принципу «Всех из всех»)

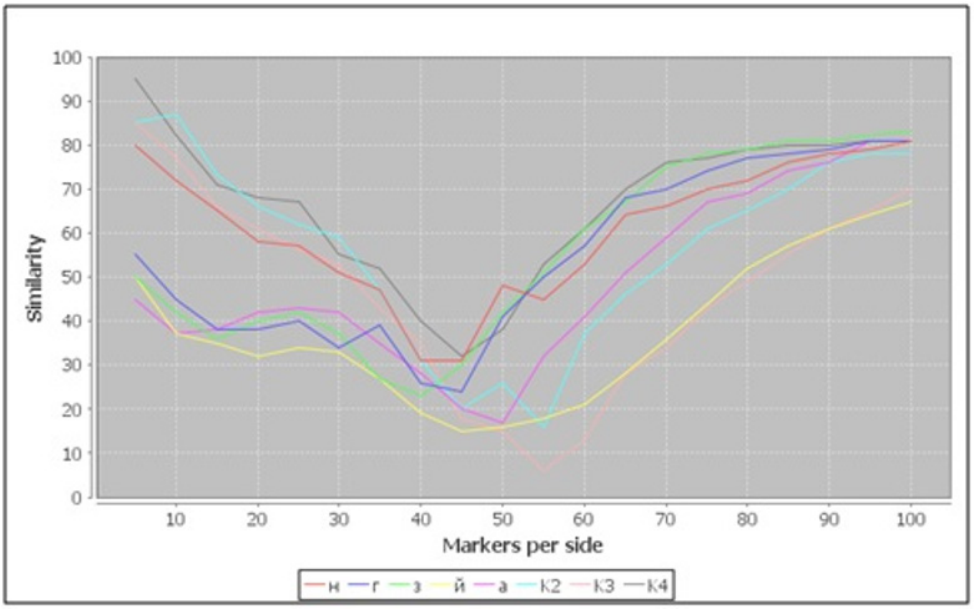

Рис. 3. График зависимости результатов сравнения от числа маркеров (метод Построение упорядоченных пар)

символа на изображении – чем меньше разрешение исходного изображения, тем раньше будет достигнуто это число маркеров. Применение же малого числа маркеров невозможно из-за риска столкновения с недостаточно четкой линией контура символа. В этом случае – некоторые из маркеров могут «пролететь» сквозь такую границу, что, при малом количестве точек, сильно изменит результаты сравнения. Таким образом, метод Построение упорядоченных пар оказался недостаточно надежным для использования.

При рассмотрении графика для метода Сравнение маркеров по принципу «Всех из всех» можно обратить внимание на следующий момент. Несмотря на стабильность результатов и прогнозируемое распределение «близости» контрольных символов к образцу, степень похожести символов, отличных от «к», кажется несколько завышенной. В попытке избавиться от постоянной составляющей в функции похожести, был применен метод «Сравнение по группам».

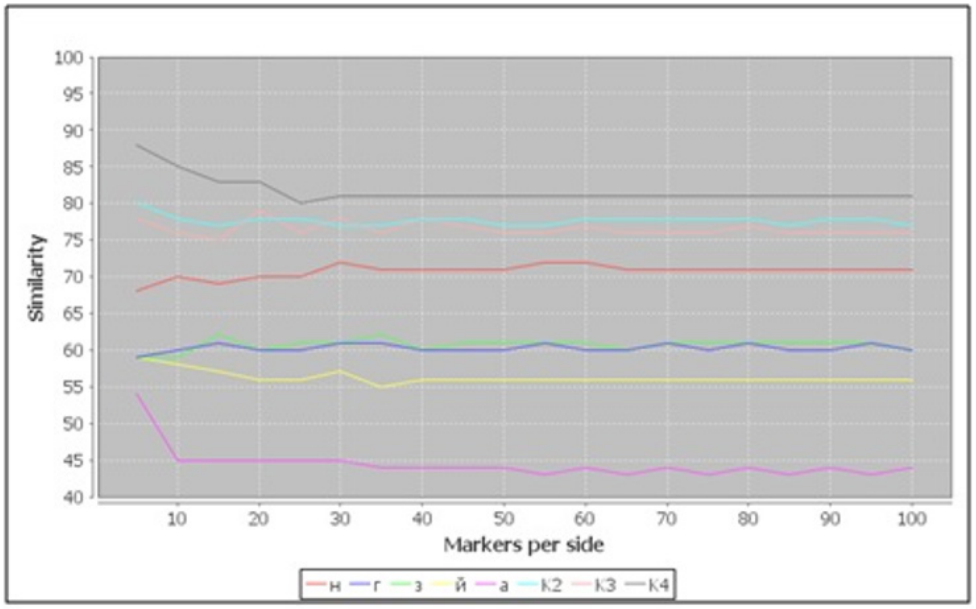

Рис. 4. График зависимости результатов сравнения от числа маркеров (метод Сравнение по группам)

Из графика на рис. 4 можно увидеть, что метод «Сравнение по группам» характеристик сохраняет стабильность и корректность результатов, при этом процент совпадения снизился за счет вывода из анализа постоянной составляющей. Таким образом, можно достичь более высокой точности результатов.

Руководствуясь подобными графиками можно выбрать оптимальное число маркеров, размещаемых с каждой стороны изображения.

ЗАКЛЮЧЕНИЕ

В работе создан программный продукт, работающий в соответствии с предложенными методами распознавания символов. На данный момент осуществляет распознавание рукописные отдельные символы русского алфавита. Как видно из представленных результатов уровень распознавания отдельных рукописных символов высокий для применения.

СПИСОК ЛИТЕРАТУРЫ

1. *Ладыженский Ю. В.* Программная реализация технологии распознавания текстовой информации / Ю. В. Ладыженский, В. В. Алейкин // Материалы V международной научно-технической конференции студентов, аспирантов и молодых ученых. – Донецк : ДоНТУ, 2009. – С. 261–264.

2. *Шапиро Л.* Компьютерное зрение / Л. Шапиро, Дж. Стокман; перевод с англ. А. Богуславский, С. Соколов. – М.: «БИНОМ Лаборатория знаний», 2009. – 760 с.

3. *Форсат Д.* Компьютерное зрение. Современный подход / Дэвид А. Форсайт, Жан Понс; перевод с англ. А. Назаренко, И. Дорошенко. – М. : «Вильямс», 2004. – 928 с.

Тимофеева Н. Е. – зав. лаборатории теоретических проблем информатики и ее приложений кафедры дискретной математики и информационных технологий, Саратовский национальный исследовательский государственный университет им. Н. Г. Чернышевского.
Тел.: +7 904 243 1815
E-mail: timofeevane@yandex.ru

Гераськин А. С. – канд. пед. наук, доцент кафедры теоретических основ компьютерной безопасности и криптографии, Саратовский национальный исследовательский государственный университет им. Н. Г. Чернышевского.
Тел.: +7 905 381 6366
E-mail: gerascinas@mail.ru

Timofeeva N. E. – Chief of Laboratory of Theoretical Problems of computer science and its applications of the Chair of Discrete Mathematics and Informational Technologies, National Research Saratov State University.
Tel.: +7 904 243 1815
E-mail: imofeevane@yandex.ru

Geraskin A. S. – Candidate of Pedagogical Sciences, Associate Professor of the Chair of Computer Science and Cryptography Theory, National Research Saratov State University.
Tel.: +7 905 381 6366
E-mail: gerascinas@mail.ru