УДК 004.912+004.932.2 **DOI** 10.15622/sp.60.5

А.В. Козачок, С.А. Копылов, Р.В. Мещеряков, О.О. Евсютин, Л.М. Туан

ПОДХОД К ИЗВЛЕЧЕНИЮ РОБАСТНОГО ВОДЯНОГО ЗНАКА ИЗ ИЗОБРАЖЕНИЙ, СОДЕРЖАЩИХ ТЕКСТ

Козачок А.В., Копылов С.А., Мещеряков Р.В., Евсютин О.О., Туан Л.М. **Подход к извлечению робастного водяного знака из изображений, содержащих текст.**

Аннотация. Представлен подход к извлечению робастного водяного знака из изображений, содержащих текст. Извлечение данных основано на разработанном подходе к внедрению в текстовые данные робастного водяного знака, отличающегося устойчивостью к преобразованию текстовых данных в формат изображения. Проведен сравнительный анализ существующих подходов к стеганографическому внедрению информации в текстовые данные, определены их достоинства и недостатки. Обоснован выбор группы методов стеганографического внедрения информации, основанных на форматировании текста. В качестве алгоритма встраивания выбран подход на основе изменения величины межстрочных интервалов. Приведены блок-схема и описание разработанного алгоритма встраивания информации в текстовые данные. Осуществлена экспериментальная оценка емкости встраивания и границ перцептивной невидимости встроенных данных. На основе существующих ограничений разработан подход к извлечению встроенной информации из изображений, содержащих робастный водяной знак. Базовым в процедуре извлечения встроенной информации выбрано преобразование Радона, позволяющее извлекать значения величин межстрочных интервалов. Для выделения значений битов встроенной информации использован подход на основе разделения смеси нормальных распределений, поскольку полученное распределение является бимодальным. Экспериментально установлены границы применимости, а также дана оценка робастности разработанного подхода встраивания к осуществлению различных преобразований. Определены следующие параметры робастности разработанного подхода к осуществлению преобразований: поворот изображения, содержащего встроенные данные на любой угол, масштабирование изображения с множителем масштабирования, не превышающим 1,5; преобразование в любой формат растрового изображения; применение медианного фильтра к изображению с пределом ядра свертки не более 9 пикселей, гауссовского фильтра размытия — с пределом показателя размытия не более 8 пикселей и усредненного фильтра с пределом ядра свертки не более 5 пикселей.

Ключевые слова: защита информации, текстовая стеганография, цифровой водяной знак, преобразование Радона, разделение смеси распределений.

1. Введение. Стремительное развитие «Интернета вещей» привело к увеличению устройств, имеющих доступ к глобальной сети [1, 2]. С ростом объемов данных и степени разнородности устройств в значительной степени возросло и количество инцидентов информационной безопасности, связанных с нарушением авторских прав владельцев текстовой информации и подменой их содержимого. По данным аналитической компании Solar Security во втором полугодии 2017 года наблюдался стабильный рост инцидентов, связанный с утечками конфиденциальных

данных и компрометацией учетных записей [3]. Одним из типов данных, подвергшихся компрометации являются текстовые данные, содержащие конфиденциальную информацию. Повышению защищенности текстовых данных, представленных в электронном виде, посвящено множество работ [4-8]. В то же время задача по защите текстовых данных, представленных в напечатанном виде, остается наиболее актуальным направлением исследований. Данная задача может быть решена посредством скрытого внедрения в текстовые данные идентифицирующей информации, которая может характеризовать владельца данных, сами данные или содержать другую метаинформацию [9].

В настоящее время для встраивания информации в текстовые данные используется технология внедрения цифровых водяных знаков (ЦВЗ). ЦВЗ могут быть использованы для защиты авторских прав, обнаружения подмены данных, проверки подлинности или вставки дополнительной информации. ЦВЗ представляет собой видимый или невидимый знак (информационную последовательность), встраиваемый в исходные данные. В аспекте работы классификация ЦВЗ рассматривается по двум направлениям (рисунок 1) [10].

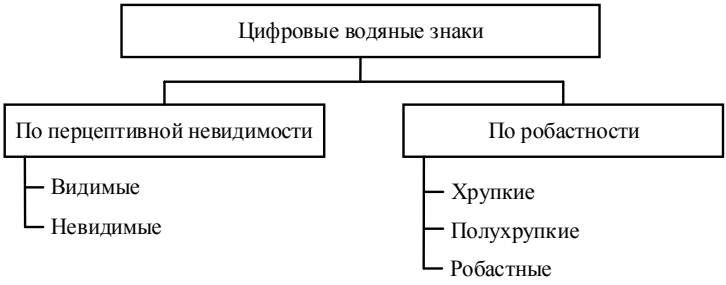


Рис. 1. Классификация ЦВЗ

Примером внедрения видимого ЦВЗ в текстовые данные является встраивание электронной подписи. Недостаток использования видимых ЦВЗ заключается в возможности удаления или модификации встроенных данных, что, в свою очередь, не позволяет правильно идентифицировать владельца либо определить подмену области исходных данных. Невидимые ЦВЗ лишены данного недостатка, однако в процессе их извлечения необходимы дополнительные этапы, направленные на определение наличия водяного знака и восстановление данных, которые они содержат.

По критерию надежности ЦВЗ могут быть представлены следующими группами:

– робастный водяной знак (РВЗ) — обеспечивает устойчивость к внесению искажений и осуществлению различных преобразований;

– хрупкий водяной знак — разрушается при внесении искажений и осуществлении различных преобразований;

– полухрупкий водяной знак — разрушается отдельным типом искажений (преобразований), однако обеспечивает устойчивость к другим типам искажений (преобразований) [11].

В случае применения преобразований и внесения искажений в хрупкие и полухрупкие водяные знаки невозможно осуществить извлечение встроенной информации ввиду разрушения водяного знака. Указанная особенность не позволяет использовать данные ЦВЗ в процессе защиты авторских прав и проверки подлинности текстовых данных. В свою очередь РВЗ способен обеспечить свойство инвариантности встроенных данных в случае применения преобразований и внесения искажений.

Невидимые робастные водяные знаки характеризуются следующими параметрами [12-14]:

– емкость встраивания (полезная нагрузка) — количество информации, которая может быть встроена (внедрено) в контейнер. В качестве контейнера могут выступать типы данных, пригодные для встраивания информации;

– невидимость (перцепционная прозрачность) — качественная характеристика, отражающая степень искажения контейнера встраиваемыми данными. Данная характеристика основана на перцептивном восприятии человека;

– необнаруживаемость (сложность обнаружения) — качественная характеристика, отражающая степень искажения статистических характеристик контейнера, не связанных с перцептивным восприятием человека;

– робастность — способность встроенных данных сохранять свойство инвариантности после осуществления различных преобразований над контейнером, подмены или удаления встроенных данных;

– извлекаемость — способность правильного извлечения встроенных данных из контейнера.

Для обеспечения невидимости и необнаруживаемости РВЗ в процессе встраивания в текстовые данные широкое распространение получили стеганографические методы, позволяющие осуществлять скрытое встраивание дополнительной информации в цифровые объекты за счет внесения некоторых изменений в элементы данных, составляющие цифровой объект [15, 16]. Для определения наиболее подходящего алгоритма

стеганографического встраивания информации в текстовые данные необходимо провести сравнительный анализ существующих методов текстовой стеганографии.

2. Обоснование выбора метода встраивания и информации в текстовые данные. Методы текстовой стеганографии принято разделять на следующие группы [17, 18]:

– методы, основанные на создании последовательности с учетом статистических характеристик исходного текста;

– лингвистические методы;

– методы, основанные на форматировании текста.

Методы текстовой стеганографии, основанные на создании последовательности с учетом статистических характеристик исходного текста, могут быть разделены на методы, генерирующие случайную последовательность символов, и методы, основанные на учете статистических характеристик исходного текста. В первом случае генерируется случайная последовательность символов, которая встраивается в текст. Недостаток данного метода заключается в отсутствии скрытости внедряемых данных по отношению к семантическому и визуальному анализу текста. Второй метод лишен данного недостатка. Он получил название текстовая мимикрия. При использовании текстовой мимикрии генерируется осмысленный текст на основе синтаксиса, описанного в контекстно-свободной грамматике, и встраивается информация за счет выбора из нее определенных фраз и слов [19].

Методы данной группы позволяют сохранить внедренную информацию как внутри электронного документа, так и в случае вывода его на печать. Данная группа методов характеризуется большой емкостью встраивания, наличием стойкости к операциям повторного набора текста и обработке текста средствами оптического распознавания символов. Недостатком данной группы методов является возможность обнаружения встроенных данных в процессе семантического анализа текста.

Лингвистические методы текстовой стеганографии для встраивания информации используют языковые свойства и структуру текста. Лингвистические методы сочетают в себе синтаксические, пунктуационные добавления (добавление знаков пунктуации в строго определенные места в тексте) и семантические методы (замена определенных слов в тексте на синонимы, использование кодовых таблиц, словарей сокращений и аббревиатур) [20-23] встраивания информации в исходный текст. К недостаткам данного метода относится небольшая емкость встраивания данных и высокая сложность восстановления.

Методы данной группы применяются для встраивания информации во все виды текстов. Данная группа методов обеспечивает стойкость к операциям повторного набора текста и обработки текста средствами оптического распознавания символов. В свою очередь данная группа, как и предыдущая, характеризуется невозможностью применения к официальным текстовым документам, подготовленным для печати и публикации.

К методам текстовой стеганографии, основанным на форматировании текста, относятся методы, осуществляющие форматирование текста и текстового пространства. Под форматированием текста и текстового пространства подразумевается изменение структуры текста, а также формы и представления шрифтов. К данной группе относятся следующие подходы:

– основанные на использовании пространства интервалов и отступов (сдвиг линии текста, сдвиг слова текста, сдвиг символа внутри слова, сдвиг абзацного отступа);

– основанные на использовании открытого пространства (внесение дополнительных пробелов);

– основанные на кодировании признаков (замена символов).

В алгоритмах встраивания, реализующих изменение положения линии текста (рисунок 2), кодирование информации осуществляется посредством вертикального сдвига линии текста на определенное расстояние [24, 25].

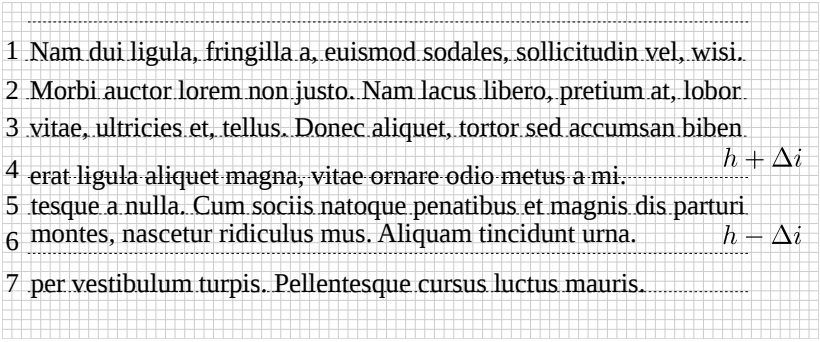


Рис. 2. Сдвиг линии текста

Алгоритмы встраивания, основанные на изменении положения слов [26], реализуют кодирование информации посредством горизонтального сдвига слов (влево или вправо на определенное расстояние) либо увеличения (уменьшения) интервала как между словами, так и между символами внутри слова.

Подход к встраиванию, основанный на сдвиге символа внутри слова, реализуется посредством кодирования информации за счет модуляции локального индекса символов текста [27] (рисунок 3). В отличие от рассмотренных ранее алгоритмов встраивания положение символа кодируется двумя битами.

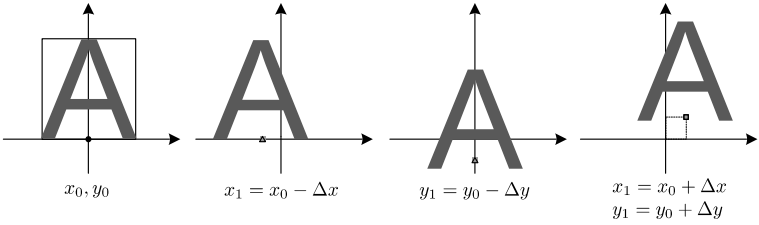


Рис. 3. Изменение положения символа

В алгоритмах встраивания, основанных на изменении значения отступа и параметров текста, кодирование информации осуществляется посредством изменения величины отступа (в том числе абзацного), табуляции, значения полей, колонтитулов и выравнивания текста.

Особенность подходов встраивания, основанных на использовании пространства интервалов и отступов, заключается в применении небольших значений величин отступа и сдвига для обеспечения перцептивной невидимости встраиваемой информации.

Вторая группа подходов встраивания, основанных на использовании открытого пространства, реализует схему внесения дополнительных пробелов в исходный текст. В качестве свободного (пустого) пространства может выступать: пустое пространство в конце каждой строки, пробелы между словами в предложении и пробелы, следующие за пунктуационными знаками [28].

Алгоритмы встраивания, основанные на внесении в текст дополнительных пробелов [12], осуществляют кодирование информации посредством сопоставления единичному дополнительному пробелу значения «1», а двойному пробелу — «0».

Алгоритмы встраивания, основанные на внедрении дополнительных пробелов в конец каждой строки, кодируют один бит встраиваемой информации двумя дополнительными пробелами в конце строки, два бита информации — четырьмя пробелами и так далее в зависимости от достижимой емкости встраивания [26].

В подходах к встраиванию, основанных на кодировании признаков, встраивание информации в исходный текст осуществляется посредством

замены одного и более символов текста. Для этого исходный текст анализируется и производится выборка символов, пригодных для встраивания. Такими символами могут быть символы, содержащие диакритические знаки, кодирование информации в которых производится путем их смещения (удаления), или символы, содержащие выносные линии, кодирование информации в которых осуществляется посредством изменения их длины и ширины [29].

Помимо изменения формы и размера символов алгоритмы встраивания, основанные на кодировании признаков, могут использовать особенности стандартов кодирования. Алгоритм встраивания, основанный на стандарте кодирования Unicode [30], осуществляет встраивание информации посредством замены одного из 13 символов английского алфавита, имеющих аналогичное представление в других алфавитах, на один из трех символов-аналогов в формате Unicode.

Кодирование признаков в алгоритмах встраивания может быть основано на сходстве некоторых типов шрифтов и реализовано посредством замены символов одного типа шрифта на другой. Для реализации алгоритма встраивания информации необходимо определить набор шрифтов схожих с исходным, составить таблицу замены и осуществить внедрение информации.

Методы текстовой стеганографии, основанные на форматировании текста, могут применяться в случаях, когда необходимо сохранение внедренной информации при печати электронного документа. Данная группа методов обеспечивает высокий уровень невидимости встраиваемых данных по отношению к визуальному и семантическому анализу текста. Недостаток, присущий всем подходам данной группы, заключается в отсутствии стойкости к операциям повторного набора текста и обработки текста средствами оптического распознавания символов. Кроме того, рассмотренные подходы характеризуются небольшой емкостью встраивания данных.

Указанные достоинства и недостатки методов текстовой стеганографии позволяют сделать вывод о применимости методов, основанных на форматировании текста, для внедрения РВЗ в текстовые данные, подготовленные для печати. Особенность текстовых данных, подготовленных к печати, заключается в невозможности изменения семантической и синтаксической структуры текста. Данная особенность накладывает ограничение на использование лингвистических методов и методов, основанных на создании последовательностей, в процессе внедрения РВЗ в текстовые данные.

Использование в процессе внедрения РВЗ подходов встраивания, основанных на внесении дополнительных пробелов и кодировании признаков в текстовые данные, подготовленные для печати, накладывает дополнительные требования на этапы обнаружения и извлечения встроенных данных. В процессе извлечения встроенных данных необходимо точно определить количество дополнительно встроенных пробелов или расположение измененных символов внутри текста, что является трудновыполнимой задачей для существующих систем оптического распознавания символов. Для решения данной задачи наиболее подходящими являются следующие подходы к встраиванию: подход, основанный на изменении величины межстрочного интервала, и подход, основанный на изменении величины межсимвольного интервала.

Подход к внедрению информации, основанный на изменении межсимвольного интервала, применим при наборе текста с использованием моноширинных шрифтов, что не соответствует существующим правилам подготовки официальных документов и документов, подготовленных к печати или опубликованию. Подход к внедрению данных, основанный на изменении межстрочного интервала (положении линий текста) лишен указанного недостатка и может быть использован в процессе стеганографического встраивания РВЗ в текстовые данные, подготовленные для печати.

Таким образом, указанные ограничения позволяют выбрать в качестве подхода к встраиванию информации в текстовые данные разработанный алгоритм стеганографического внедрения РВЗ, основанный на изменении величины межстрочного интервала.

3. Подход к встраиванию информации в текстовые данные. Вариант встраивания информации на основе изменения межстрочного интервала (рисунок 2) позволяет кодировать 1 бит информации с помощью изменения межстрочного интервала текста на величину Δ. В разработанном алгоритме встраивание информации реализуется следующим образом:

– увеличение величины межстрочного интервала на установленное значение Δ между соседними строками текста интерпретируется как значение «1»;

– отсутствие изменений в величине межстрочного интервала между соседними строками текста интерпретируется как значение «0».

Блок-схема алгоритма стеганографического внедрения информации в текстовые данные представлена в листинге 1.

Исходными данными для работы алгоритма является текстовый документ TD_0 (документ-контейнер), содержащий текстовые данные.

Data: Текстовый документ TD_0, встраиваемая информация I
Result: Подписанный текстовый документ TD_s
1 $Len \leftarrow$ **GetLength** *(I)*
2 $N \leftarrow$ **CountLines** *(TD_0)*
3 **if** $N > (Len + 1)$ **then**
4 **for** $i \leftarrow 0$ **to** $(N - 2)$ **do**
5 $j \leftarrow i \bmod Len$
6 **if** $I_j = 1$ **then**
7 $TD_0 \leftarrow$ **Embed** *(TD_0,i)*
8 $TD_s \leftarrow TD_0$
9 **return** TD_s

Листинг 1. Алгоритм внедрения информации в текстовый документ

Допускается использование следующих форматов текстовых документов:

– RTF (Rich Text Format) — спецификация [31];

– PDF (Portable Document Format) — стандарт [32, 33].

Первым этапом работы алгоритма стеганографического внедрения информации (шаги 1-2) является определение размера встраиваемой информации и достижимой емкости встраивания функцией $GetLength$ и $CountLines$ соответственно.

На втором этапе (шаг 3) определяется возможность встраивания информации в исходные текстовые данные. На этапе кодирования встраиваемой информации (шаги 4-7) кодовая последовательность встраивается в исходный текстовый документ следующим образом:

– внедрение информации осуществляется с первой (i) до ($N - 2$) строки, где N определяет достижимую емкость встраивания;

– в i строке считывается символ кодовой комбинации I;

– если считанный символ соответствует «1», то межстрочный интервал между i и $i + 1$ строками увеличивается на величину Δ.

По окончании работы алгоритма стеганографического внедрения информации формируется подписанный текстовый документ TD_s, содержащий встраиваемую информацию I.

Для определения возможности реализации указанного подхода встраивания, достижимой емкости встраивания и установления границ перцептивной невидимости встроенной информации в текстовые данные была проведена их экспериментальная оценка разработанного алгоритма.

4. Оценка емкости встраивания и перцептивной невидимости подхода на основе изменения межстрочного интервала. Проведенная группа экспериментов позволила определить величину предельно до-

стижимой емкости встраивания, установить зависимость достижимой емкости встраивания от размера и гарнитуры используемого шрифта и изменения величины межстрочного интервала. Кроме того, проведена оценка перцептивной невидимости встроенных данных и определены пороговые значения изменения величины межстрочного интервала.

Исходными данными являлся текст, созданный в системе компьютерной верстки LaTeX, размером не менее 10 страниц со следующими размерами полей:

– верхнее, нижнее — 20 мм;
– левое — 30 мм;
– правое — 10 мм.

Предельно достижимая емкость встраивания алгоритма характеризуется количеством межстрочных интервалов текста на одной странице. Для установления зависимости между емкостью встраивания, параметрами шрифта и величиной межстрочного интервала были выбраны следующие варьируемые параметры:

– межстрочный интервал: одинарный, полуторный, двойной.
– кегль шрифта: 10 пт, 12 пт и 14 пт.
– гарнитура используемого шрифта:

1. Шрифт с засечками (serif) — Computer Modern Roman (аналог Times New Roman).

2. Шрифт без засечек (sans serif) — Computer Modern Sans Serif (аналог Arial).

3. Моноширинный шрифт (monospace) — Computer Modern Typewriter (аналог Courier New).

Полученные значения предельно достижимой емкости встраивания представлены в таблице 1.

Таблица 1. Зависимость емкости встраивания от изменения межстрочного интервала и кегля шрифта

Кегль шрифта	Величина межстрочного интервала	Предельно достижимая емкость встраивания
10	1	60
10	1,5	40
10	2	30
12	1	49
12	1,5	33
12	2	24
14	1	42
14	1,5	28
14	2	21

Анализ полученных результатов позволяет сделать вывод о зависимости емкости встраивания от величины межстрочного интервала и размера кегля шрифта, а также об отсутствии зависимости от гарнитуры используемого шрифта. Максимально достижимое значение емкости встраивания составляет 60 бит (при использовании шрифта с кеглем 10 пт и величине межстрочного интервала 1), минимальное — 21 бит (при использовании шрифта с кеглем 14 пт и величине межстрочного интервала 2).

В ходе проведения второй группы экспериментов была произведена оценка перцептивной невидимости встраиваемых данных. Кодирование информации реализовано следующим образом:

– величина межстрочного интервала после каждой нечетной строки текста не изменяется и интерпретируется как значение «0»;

– величина межстрочного интервала после каждой четной строки текста увеличивается на установленное значение и интерпретируется как значение «1».

В данном эксперименте использовался один абзац текста с гарнитурой Computer Modern Roman и кеглем 14 пт. Для определения порога перцептивной невидимости использовались следующие варьируемые параметры:

1. Шрифт Computer Modern Roman 14 пт (4,94 мм) с межстрочным интервалом, равным 1 (4,94 мм):

– увеличение межстрочного интервала до 1,25 (6,17 мм) с шагом 0,05 (0,245 мм);

– уменьшение межстрочного интервала до 0,75 (3,71 мм) с шагом 0,05 (0,245 мм).

2. Шрифт Computer Modern Roman 14 пт (4,94 мм) с межстрочным интервалом, равным 1,5 (7,41 мм):

– увеличение межстрочного интервала до 1,25 (8,64 мм) с шагом 0,05 (0,245 мм);

– уменьшение межстрочного интервала до 0,75 (6,17 мм) с шагом 0,05 (0,245 мм).

В результате визуального анализа полученных данных экспертным путем был установлен порог перцептивной невидимости встроенных данных:

– 0,85-1,15 (4,20-5,68 мм) — для одинарного межстрочного интервала;

– 1,35-1,65 (6,67-8,15 мм) — для полуторного межстрочного интервала.

Полученные результаты емкости встраивания и перцептивной невидимости позволяют использовать разработанный подход к внедрению РВЗ в текстовые данные для решения задачи по защите текстовых данных, подготовленных к печати.

Процесс извлечения встроенных данных накладывает дополнительные требования по точности обнаружения встроенного РВЗ и определения величин межстрочных интервалов. Кроме того, извлечение встроенного РВЗ происходит из изображения, содержащего текст, что требует введения в процесс извлечения этапа предварительной обработки изображения.

5. Подход к извлечению информации из изображений со встроенным РВЗ. Для реализации указанных требований и обеспечения достоверности извлечения встроенных данных был разработан подход к извлечению информации из изображений со встроенным РВЗ, состоящий из следующих этапов:

– предварительная обработка изображения;
– обнаружение линий текста;
– определение массива значений межстрочных интервалов;
– декодирование встроенной информации.

Отличительной особенностью разработанного подхода является применение нормального преобразования Радона для обнаружения линий текста.

5.1. Преобразование Радона. Преобразование Радона $\check{g}(p,\tau)$ непрерывной двумерной функции $f(x,y)$ определяется посредством суммирования (интегрирования) значений f вдоль наклонных линий [34]. Положение линии определяется параметрами наклона p и смещения линии τ:

$$\check{g}(p,\tau) = \int\limits_{-\infty}^{\infty} f(x, p\cdot x + \tau)\, dx. \qquad (1)$$

К свойствам преобразования Радона относятся:

– обратимость — возможность восстановить исходную функцию по ее преобразованию;

– линейность — преобразование Радона взвешенной суммы функций равно взвешенной сумме отдельно преобразованных функции;

– смещение функции — наклон линии не может быть изменен посредством применения преобразования;

– масштабируемость.

Выражение (1) представляет собой одну из нескольких форм записи преобразования. В общем виде преобразование Радона описывается следующим образом:

$$\check{g}(\xi_0, \xi_1, \xi_2) = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f(x, y)\, \delta(\xi_0 - \xi_1 \cdot x - \xi_2 \cdot x)\, dx\, dy. \qquad (2)$$

Особенность прямого преобразования Радона заключается в преобразовании каждой линии изображения в определенным образом расположенные пики, соответствующие параметрам этих линий. Обратное преобразование Радона позволяет восстановить из полученных пиков исходные линии. Из свойства обратимости преобразования Радона следует, что преобразование Радона, примененное к точке, трансформирует ее в линию, а примененное к линии — в точку.

Если базовая функция $f(x, y)$ не имеет предпочтительной ориентации, то уравнение прямой может быть представлено в нормальном виде:

$$\rho = x \cdot \cos\theta + y \cdot \sin\theta. \qquad (3)$$

Подставляя полученные значения $(\xi_0, \xi_1, \xi_2) = (\rho, \cos\theta, \sin\theta)$ в выражение (2), получим нормальное преобразование Радона:

$$\check{g}(\rho, \theta) = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} f(x, y)\, \delta(\rho - x \cdot \cos\theta - y \cdot \sin\theta)\, dx\, dy. \qquad (4)$$

Нормальное преобразование Радона имеет простой геометрический смысл — это интеграл от функции $f(x, y)$ вдоль прямой, описанной выражением (3) и повернутой на угол θ, перпендикулярной вектору $(\cos\theta, \sin\theta)$, проходящей на расстоянии ρ от начала координат [35].

В случае применения нормального преобразования Радона $R_\theta(x')$ к изображению, описанному функцией $f(x, y)$, результат преобразования представляет собой набор линейных интегралов от функции $f(x, y)$, полученный при вращении оси абсцисс (Рисунок 4). Другими словами, для представления изображения через нормальное преобразование Радона необходимо взять несколько параллельно-лучевых отображений изображения, вычисленных путем поворота исходной линии (оси абсцисс) вокруг центра изображения на различные углы [36].

Рассмотренное преобразование применяется в разработанном алгоритме извлечения встроенных данных из изображений, содержащих текст со встроенным РВЗ, на этапе обнаружения линий текста. Также

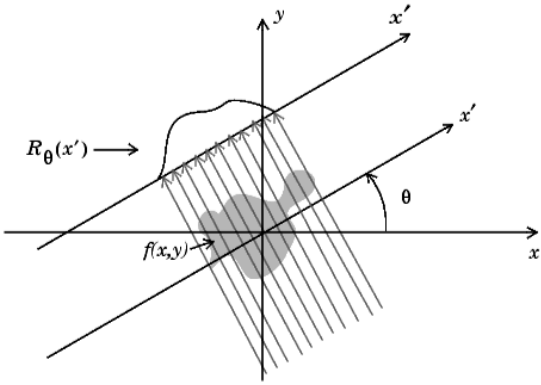


Рис. 4. Графическая интерпретация преобразования Радона

данное преобразование позволяет определить угол поворота изображения, содержащего текст, и учесть его при извлечении линий текста.

5.2. Алгоритм извлечения встроенной информации из изображений. Алгоритм извлечения информации, встроенной в текстовые данные из изображений, содержащих робастный водяной знак, представлен в листинге 2.

В качестве исходных данных выступает изображение, содержащее встроенную информацию в текстовые данные посредством алгоритма стеганографического внедрения, основанного на изменении межстрочного интервала (листинг 1).

На первом этапе осуществляется предварительная обработка исходного изображения Im_t (шаги 1-2). Функция $ConvertToGray$ преобразует цветовое пространство изображения Im_t к полутоновому изображению (в градациях серого) Im_{grey}. Компонента яркости полутонового изображения Im_{grey} подвергается среднеарифметической фильтрации (функция $Filtration$). В результате обработки исходного изображения формируется нормализованное изображение в градациях серого Im_{filt}.

В процессе обнаружения линий текста (шаг 3) нормализованное изображение Im_{filt} преобразуется в синограмму $sinogram$ посредством нормального преобразования Радона. Под синограммой понимается двумерное распределение одномерных проекций слоя объекта как функции угла проекции, где угол проекции расположен по оси ординат, а линей-

Data: Изображение, содержащее текст Im_t
Result: Встроенная информация I_e
1 $Im_{grey} \leftarrow$ **ConvertToGray** *(Im_t)*
2 $Im_{filt} \leftarrow$ **Filtration** *(Im_{grey})*
3 $sinogram \leftarrow$ **RadonTransform** *(Im_{filt})*
4 **for** $l \leftarrow 0$ **to** 180 **do**
5 $\quad R[l] \leftarrow$ **RmsCalculation** *($sinogram, l$)*
6 $rot \leftarrow$ **Argmax***(R)*
7 $row \leftarrow sinogram[rot]$
8 $M \leftarrow$ **FindPicks***(row)*
9 $min\_mode, max\_mode \leftarrow$ **FindModes***(M)*
10 $D \leftarrow$ **CorrectErrors***(M)*
11 **if Std***(D)* $> 0,8$ **then**
12 $\quad B \leftarrow$ **GaussianMixtureModel***(D, min_mode, max_mode)*
13 **else**
14 $\quad b_i \leftarrow 0,\ i = \overline{1, |D|}$
15 $\quad B \leftarrow \{b_i\}$
16 $I_e \leftarrow B$
17 **return** I_e

Листинг 2. Извлечение встроенной информации из изображения

ная проекция расположена по оси абсцисс [37]. Данное преобразование осуществляется функцией $RadonTransform$.

Из сформированной синограммы посредством функции $RmsCalculation$ для каждого ряда (угла поворота) вычисляется среднеквадратичное значение ($R[l]$) согласно выражения (шаги 4-5):

$$Rms = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}},$$

где $x_1, x_2, \ldots, x_n$ — значения синограммы по углу l.

Затем из синограммы посредством функции $Argmax$ выбирается ряд row, обладающий наиболее выровненным по линии чередованием белых полос и черных точек. Данный ряд определяет положение текста и угол поворота rot исходного изображения.

Этап вычисления величин межстрочных интервалов текста изображения состоит из обнаружения линий текста и вычисления величин межстрочных значений функцией $FindPicks$. На первом этапе происходит обнаружение линий текста. Под линией текста понимается виртуальная линия, по которой выровнен или объединен текст [36]. Для извлечения линий текста в row осуществляется поиск индексов смены знака, которые соответствуют линиям исходного текста. Расстояние между индексами смены знака определяет величину межстрочного интервала между двумя

строками исходного текста. Массив величин межстрочных интервалов M вычисляется как разность между двумя индексами смены знака.

На этапе коррекции ошибок (шаг 10) полученные значения M проверяются на наличие ошибок первого и второго рода. Под ошибками первого рода понимается извлечение межстрочных интервалов, не присутствующих в исходном тексте. Ошибки второго рода характеризуются пропуском межстрочных интервалов в исходном тексте. Для обнаружения ошибок функцией $FindModes$ вычисляется верхняя и нижняя моды min_mode, max_mode массива M. Если элемент массива больше, чем значение $1,7 \cdot max\_mode$, принимается решение о наличии ошибки второго рода. Для исправления обнаруженной ошибки указанный элемент массива заменяется на два элемента, каждый из которых имеет значение max_mode. Ошибкам первого рода соответствуют величины межстрочных интервалов массива M меньшие, чем $0,6 \cdot min\_mode$. Исправление ошибок первого рода осуществляется путем удаления данного элемента из массива M.

На шаге 11-15 осуществляется преобразование скорректированного массива D в двоичный вид. Полученный массив D можно интерпретировать как бимодальное распределение (значений нуля и единицы). Таким образом, задача преобразования полученной последовательности данных в двоичный вид может быть решена посредством применения методов математической статистики, направленных на разделение смеси распределений. В качестве метода разделения смеси используется модель разделения смеси нормальных распределений [38].

5.3. Разделение смеси распределений. Основная задача разделения заключается в восстановлении компонентов $f_\omega(X;\theta(\omega))$ и смешивающей функции $P(\omega)$ по заданной левой части $f(X)$ соотношения:

$$f(X) = \int_{-\infty}^{\omega} f_\omega(X;\theta(\omega))\, dP(\omega) \tag{5}$$

и называется задачей идентификации компонентов смеси [39].

Выражение (5) представляет собой функцию плотности (полигон вероятностей) распределения называемой смесью (или P-смесью) распределения семейства F.

Данное разделение может быть реализовано в случае, если $P = \{P(\omega)\}$ представляет собой семейство смешивающих функций распределения и имеется двухпараметрическое семейство p-мерных плотностей (полигонов вероятностей) распределения:

$$F = \{f_\omega(X; \theta(\omega))\}, \tag{6}$$

где одномерный (целочисленный или непрерывный) параметр ω в качестве нижнего индекса функции f определяет специфику общего вида каждого компонента — распределения смеси, а в качестве аргумента при многомерном параметре θ определяет зависимость значений хотя бы части компонентов этого параметра от того, в каком именно составляющем распределении f_ω он присутствует.

Тогда задача разделения смеси распределений заключается в выборке классифицируемых наблюдений $X_1, X_2, \ldots, X_n$, извлеченной из генеральной совокупности, являющейся смесью, описываемой выражением 5, генеральных совокупностей типа, приведенных в выражении 6.

Если в качестве функции плотности вероятности выступает плотность вероятности нормального закона распределения:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\dfrac{(x-m)^2}{2\sigma^2}},$$

то данный метод разделения смесей распределений называется гауссовской моделью разделения смеси распределения.

Преобразование массива D в двоичный массив B зависит от величины среднеквадратического отклонения значений элементов массива D. Функция Std осуществляет вычисление среднеквадратического отклонения значений элементов массива D. Если полученное значение среднеквадратического отклонения превышает 0,8, то бинаризация осуществляется с использованием функции $GaussianMixtureModel$ (шаг 12), реализующей модель разделения смеси нормальных распределений. Интерпретация массива D как дискретного бимодального распределения позволяет разделить его на два класса. Класс, характеризующийся меньшим значением моды элементов, соответствует «0», а класс, характеризующийся большим значение моды, соответствует «1» в двоичном массиве B. При значении $Std(D) < 0,8$ массив D считается унимодальным и принимается решение об отсутствии встроенной информации. В результате чего на шагах 14-15 формируется двоичный массив B, состоящий из «0» элементов. Полученный двоичный массив B представляет собой встроенные данные I_e.

По окончании работы алгоритма извлечения информации, встроенной в изображение, содержащее робастный водяной знак, делается вывод

о подлинности полученных данных и соответствии их подписанному владельцу.

Для разработанного алгоритма извлечения информации, встроенной в текстовые данные из изображения, была произведена оценка следующих основных свойств:

– результативность;
– элементарность;
– корректность;
– вычислительная сложность.

Результативность (отсутствие аварийного останова) алгоритма заключается в том, что исполнение алгоритма завершится за конечное число шагов.

Данный алгоритм является элементарным, так как содержит блоки, выполняющие простые операции.

Доказательство корректности алгоритма определяется следующим образом: для заданных параметров алгоритм всегда позволяет определить информацию, представленную РВЗ, внедренную в изображение.

Временную сложность алгоритма определим для наиболее вычислительно сложных процедур:

– преобразования Радона, имеющего сложность $T_1(180 \cdot x \cdot y)$, где x, y — размер изображения в пикселях;

– разделения смеси нормальных распределений $T_2(n \cdot k \cdot p \cdot i)$, где n — число элементов в массиве, k — число модальностей ($k = 2$), p — размерность пространства ($p = 1$), i — число итераций алгоритма.

Таким образом, общую временную сложность алгоритма можно представить как $T(180 \cdot x \cdot y + 2 \cdot n \cdot i)$.

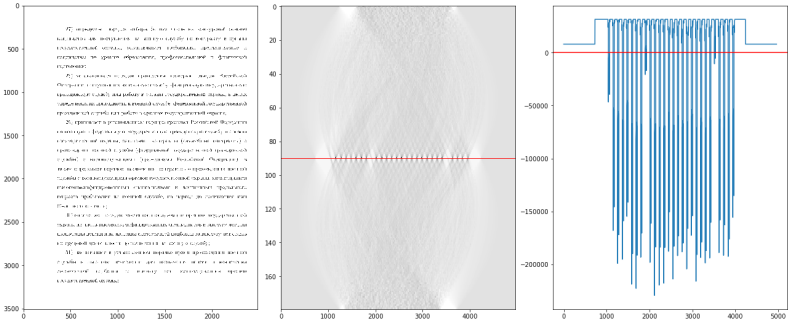
Сложность алгоритма по памяти равна $O(x \cdot y)$. Такая оценка сложности является линейной.

Для определения границ извлекаемости встроенных данных и робастности разработанного подхода к стеганографическому внедрению информации в текстовые данные РВЗ была проведена экспериментальная оценка.

6. Экспериментальная оценка применимости и робастности разработанного алгоритма извлечения встроенной информации из изображений. В ходе проведения эксперимента кодирование информации производись путем увеличения межстрочного интервала на значения, находящиеся в пределах перцептивной невидимости (от 0,01 до 0,1). Внедрение информации осуществлялось в текстовые данные, набранные шрифтом Computer Modern Roman с размером кегля 14 пт.

РВЗ представлен последовательностью «0101», циклически встроенной в исходный текст.

Результат извлечения РВЗ из изображения, содержащего встроенные данные, представлен на рисунке 5 (Рисунок 5а — исходное изображение, содержащее РВЗ; Рисунок 5б — сформированная синограмма; Рисунок 5в — результат преобразования Радона, где пики соответствуют линиям текста на изображении.



а б в

Рис. 5. Пример извлечения линий текста из изображения, содержащего текст

Полученные значения количества строк и соответствующий им массив величин межстрочных интервалов представлены в таблице 2.

Таблица 2. Массив величин межстрочных интервалов

Число строк в исходном документе	Вычисленное число строк	Массив извлеченных величин межстрочных интервалов
28	30	[83, 80, 83, 80, 82, 80, 83, 10, 4, 2, 80, 83, 80, 82, 80, 83, 80, 82, 80, 83, 80, 83, 188, 80, 83, 80, 83, 79, 83]
28	31	[85, 80, 86, 80, 85, 80, 86, 80, 12, 8, 86, 79, 86, 80, 86, 79, 86, 80, 86, 79, 86, 80, 86, 79, 86, 24, 80, 86, 80, 85]
28	27	[80, 88, 80, 89, 80, 88, 80, 89, 79, 89, 80, 88, 194, 80, 88, 80, 89, 79, 89, 80, 88, 80, 89, 80, 88, 80]

Результаты процесса исправления ошибок межстрочных интервалов приведены в таблице 3.

Результаты бинаризации извлеченных данных из изображений со встроенным РВЗ представлены в таблице 4.

Точность извлечения встроенных данных из изображений, содержащих РВЗ, представлена в таблице 5.

Таблица 3. Скорректированные массивы величин межстрочных интервалов

Число строк в исходном документе	Исправленное число строк	Массив исправленных величин межстрочных интервалов
28	28	[83, 80, 83, 80, 82, 80, 83, 80, 83, 80, 82, 80, 83, 80, 82, 80, 83, 80, 83, 83, 83, 80, 83, 80, 83, 79, 83]
28	28	[85, 80, 86, 80, 85, 80, 86, 80, 86, 79, 86, 80, 86, 79, 86, 80, 86, 79, 86, 80, 86, 79, 86, 80, 86, 80, 85]
28	28	[80, 88, 80, 89, 80, 88, 80, 89, 79, 89, 80, 88, 88, 88, 80, 88, 80, 89, 79, 89, 80, 88, 80, 89, 80, 88, 80]

Таблица 4. Результат бинаризации извлеченных данных

Встроенная информация	Двоичный массив межстрочных интервалов
1010	[1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, **1**, 1, 0, 1, 0, 1, 0, 1]
1010	[1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1]
0101	[0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, **1**, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0]

Таблица 5. Точность извлечения данных

Размер кегля (пт)	Величина изменения межстрочного интервала (множитель)	Точность извлечения (%)	Ложные срабатывания (%)	Пропуск цели (%)
14	1,01	67,8	28,6	3,6
14	1,02	60,5	16,1	23,4
14	1,03	72,9	8,6	18,5
14	1,04	95,1	1,2	3,7
14	1,05	96,3	2,5	1,2
14	1,06	96,3	2,5	1,2
14	1,07	98,8	0	1,2
14	1,08	98,8	1,2	0
14	1,09	98,8	1,2	0
14	1,10	98,8	0	1,2

Полученные значения точности извлечения данных из изображений, содержащих РВЗ, позволяют сделать вывод о зависимости результата извлечения встроенной информации от величины изменения межстрочного интервала, используемого в процессе встраивания. Так, изменение величины межстрочного интервала в диапазоне значений от 0,01 (0,05 мм) до 0,03 (0,15 мм) интервала не позволяет правильно извлечь встроенную информации ввиду высокого процента ошибок. В то же время изменение величины межстрочного интервала в диапазоне от 0,04 (0,20 мм) до 0,10 (0,50 мм) интервала характеризуются одиночными ошибками извлечения. Для исправления одиночных ошибок в процессе встраивания могут быть применены методы помехоустойчивого кодирования, что поз-

волит повысить точность извлечения встроенных данных из изображений, содержащих РВЗ.

Помимо оценки границ применимости разработанного алгоритма была произведена оценка устойчивости изображений со встроенным РВЗ к различным преобразованиям над изображениями.

Оценка стойкости разработанного алгоритма извлечения информации из изображений со встроенным РВЗ состоит в возможности достоверного извлечения информации из изображений после применения к ним следующих преобразований:

– поворот изображения;
– масштабирование;
– преобразование формата (JPEG, PNG, BMP, PDF);
– медианная фильтрация;
– гауссовская фильтрация;
– усредненная фильтрация.

Полученные результаты экспериментальной оценки стойкости разработанного алгоритма извлечения информации из изображения со встроенным РВЗ представлены в таблице 6. Пример результата извлечения данных из изображения, содержащего РВЗ, после применения гауссовского фильтра размытия с пределом показателя размытия 8 пикселей представлен на рисунке 6.

Таблица 6. Стойкость разработанного алгоритма к преобразованиям применяемым к изображениям

Тип преобразования	Устойчивость разработанного алгоритма к преобразованию
Поворот	поворот на любой угол
Масштабирование	множитель масштабирования до 1,5
Преобразование формата	JPEG, PNG, GIF, BMP, PDF
Медианная фильтрация	с пределом ядра свертки 9 пикселей
Гауссовская фильтрация	с пределом показателя размытия 8 пикселей
Усредненная фильтрация	с пределом ядра свертки 5 пикселей

7. Заключение. Разработанный подход к внедрению робастного водяного знака в текстовые данные позволяет обеспечить защиту авторских прав владельцев данных посредством внедрения в качестве РВЗ информации, идентифицирующей владельца данных, а также другой метаинформации. Кроме того, разработанный подход к внедрению робастного водяного знака обеспечивает инвариантность к рассмотренным преобразованиям и позволяет достоверно извлекать встроенные данные. Повышение точности извлечения, а также снижение количества ошибок встраивания являются направлением дальнейших исследований.

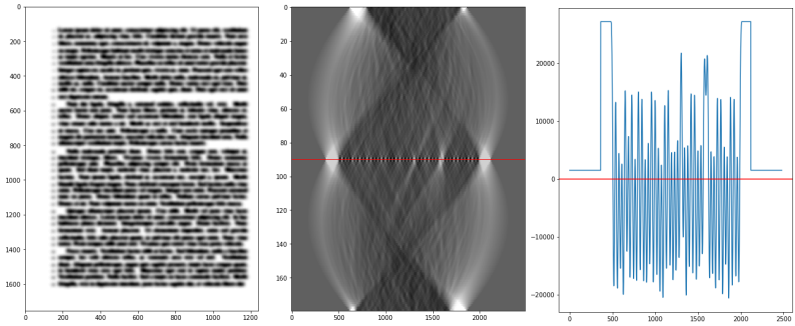


Рис. 6. Пример извлечения линий текста из изображения, содержащего текст

Литература

1. *Weber R.H.* Internet of Things – New security and privacy challenges // Computer Law & Security Review. 2010. vol. 26. no. 1. pp. 23–30.
2. *Chaudhuri A.* Internet of things data protection and privacy in the era of the General Data Protection Regulation // Journal of Data Protection & Privacy. 2016. vol. 1. no. 1. pp. 64–75.
3. JSOC Security flash report второе полугодие 2017 года. URL: solarsecurity.ru/upload/iblock/54f/flash_report_Q2_2017_041.pdf (дата обращения: 02.05.2018).
4. *Ramapriya B.* An Improved Approach of Text Steganography in Application with Rotational Symmetry // International Journal of Innovative Research in Computer and Communication Engineering. 2017. vol. 5. no. 7. pp. 12939–12947.
5. *Elmahi M.Y., Kosti Su., Sayed M.H.* Text Steganography Using Compression and Random Number Generators // International Journal of Computer Applications Technology and Research. 2017. vol. 6. no. 6. pp. 259–263.
6. *Kingslin S., Kavitha N.* Evaluative approach towards text steganographic techniques // Indian Journal of Science and Technology. 2015. vol. 8. no. 29. pp. 1–8.
7. *Odeh A., Elleithy K., Faezipour M., Abdelfattah E.* Highly efficient novel text steganography algorithmss // Systems, Applications and Technology Conference (LISAT). 2015. pp. 1–7.
8. *Rathore A.S., Rawat S.S.* A Secure Image and Text Steganography Technique // International Journal for Research in Applied Science & Engineering Technology (IJRASET). 2015. vol. 3. no. 5. pp. 506–509.
9. *Резанова З.И., Романов А.С., Мещеряков Р.В.* О выборе признаков текста, релевантных в авторoведческой экспертной деятельности // Вестник Томского государственного университета. Филология. 2013. Т. 26. №. 6. С. 38–52.
10. *Грибунин В.Г., Оков И.Н., Туринцев И.В.* Цифровая стеганография // М.: СОЛОН-Пресс. 2017. 262 с.
11. *Козачок А.В., Копылов С.А., Мещеряков Р.В., Евсютин О.О.* Анализ текущего состояния научных исследований в области робастного хэширования изображений // Компьютерная оптика. 2017. Т. 41. №. 5. С. 743–755.
12. *Salomon D.* Data privacy and security: encryption and information hiding // Springer Science & Business Media. 2003. pp. 469.
13. *Woo C.-S.* Digital image watermarking methods for copyright protection and authentication // Ph.D. Thesis. Queensland University of Technology. 2007. 223 p.

14. *Phadikar A.* Robust Watermarking Techniques for Color Images // 2009.
15. *Евсютин О.О., Шелупанов А.А., Мещеряков Р.В., Бондаренко Д.О.* Алгоритм встраивания информации в сжатые цифровые изображения на основе операции замены с применением оптимизации // Компьютерная оптика. 2017. Т. 41. №. 3. С. 412–421.
16. *Ватаманюк И.В,. Ронжин А.Л.* Применение методов оценивания размытости цифровых изображений в задаче аудиовизуального мониторинга // Информационно-управляющие системы. 2014. Т. 71. №. 4. С. 16–23.
17. *Agarwal M.* Text steganographic approaches: a comparison // 2013. arXiv preprint arXiv:1302.2718. 16 p.
18. *Pamulaparty L., Rao N.M.* Text Steganography: Review // International Journal of Computer Science and Information Technology & Security (IJCSITS). 2016. vol. 6. no. 4. pp. 80–83.
19. *Ярмолик С.В., Листопад Ю.Н.* Стеганографические методы защиты информации // Информатизация образования. 2005. №. 5. С. 64–74.
20. *Shirali-Shahreza M., Shirali-Shahreza M.H.* Text steganography in SMS // Proceedings of IEEE International Conference on Convergence Information Technology. 2007. pp. 2260–2265.
21. *Бондарчук С.С., Давыдова Е.М., Костюченко Е.Ю.* Встраивание цифровых знаков для обеспечения защиты информации // Доклады Томского государственного университета систем управления и радиоэлектроники. 2011. Т. 24. №. 2. С. 228–235.
22. *Govada S.R., Kumar B.S., Devarakonda M., Stephen M.J.* Text steganography with multi level shielding // IJCSI International Journal of Computer Science Issues. 2012. vol. 9. no. 5. pp. 401–405.
23. *Nagarhalli T.P., Bakal J.W., Jain N.*A Survey of Hindi Text Steganography // International Journal of Scientific & Engineering Research. 2016. vol. 7. no. 3. pp. 55–61.
24. *Low S.H., Maxemchuk N.F., Brassil JT., O'Gorman L.* Document marking and identification using both line and word shifting // Proceedings of IEEE INFOCOM'95 on Fourteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Bringing Information to People. 1995. vol. 2. pp. 853–860.
25. *Saraswathi V., Kingslin M.S.* Different Approaches to Text Steganography: A Comparison // International Journal of Emerging Research in Management & Technology. 2014. vol. 9359. no. 11. pp. 124–127.
26. *Por L.Y., Delina B.* Information hiding: A new approach in text steganography // Proceedings of WSEAS International Conference on Mathematics and Computers in Science and Engineering. 2008. vol. 7. pp. 689–695.
27. *Villán R. et al.* Tamper-proofing of electronic and printed text documents via robust hashing and data-hiding // Security, Steganography, and Watermarking of Multimedia Contents. 2007. pp. 6505–6517.
28. *Kumar K.A., Pabboju S., Desai N.M.S.* Advance text steganography algorithms: an overview // International Journal of Research and Applications. 2014. vol. 1. no. 1. pp. 31–35.
29. *Bhaya W., Rahma A.M., Al-nasrawi D.* Text steganography based on font type in MS-Word documents // Journal of Computer Science. 2013. vol. 9. no. 7. pp. 898–904.
30. *Rahma A.M.S., Bhaya W.S., Al-Nasrawi D.A.* Text steganography based on unicode of characters in multilingual // International Journal of Engineering Research and Applications. 2013. vol. 3. no. 4. pp. 1153–1165.
31. Word 2007: Rich Text Format (RTF) Specification, version 1.9.1. URL: www.microsoft.com/en-us/download/details.aspx?id=10725 (дата обращения: 02.05.2018).

32. 171/SC 2: ISO 32000–1: 2008 Document Management-Portable Document Format-Part 1 // ISO. TC. 2008. 747 p.
33. 171/SC 2: ISO 32000–2: 2017 Document Management-Portable Document Format-Part 2 // ISO. TC. 2017. 371 p.
34. *Toft P.A., Sørensen J.A.* The Radon transform-theory and implementation // Ph.D. Thesis. Technical University of Denmark. 1996. 321 p.
35. *Волков В.А.* Ряды Фурье. Интегральные преобразования Фурье и Радона: учебно-методическое пособие // Издательство Уральского университета. 2014. 32 с.
36. *Bahaghighat M.K., Mohammadi J.* Novel approach for baseline detection and Text line segmentation // International Journal of Computer Applications. 2012. vol. 51. no. 2. pp. 9–16.
37. ГОСТ Р МЭК/ТО 61948-2–2008 Оборудование для радионуклидной диагностики. Эксплуатационные испытания. Часть 2. Сцинтилляционные камеры и однофотонные компьютерные томографы // М.: Госстандарт России. 2009. 8 с.
38. *Reynolds D.* Gaussian mixture models // Encyclopedia of biometrics. 2015. pp. 827–832.
39. *Айвазян С.А., Бухштабер В.М., Енюков Е.С.* Прикладная статистика: Классификация и снижение размерности // М.: Финансы и статистика. 1989. 607 с.

Козачок Александр Васильевич — к-т техн. наук, сотрудник, Академия Федеральной службы охраны Российской Федерации. Область научных интересов: информационная безопасность, защита от несанкционированного доступа, математическая криптография, теоретические проблемы информатики. Число научных публикаций — 105. alex.totrin@gmail.com; Приборостроительная, 35, Орел, 302034; р.т.: +7(486) 254-99-33.

Копылов Сергей Александрович — сотрудник, Академия Федеральной службы охраны Российской Федерации. Область научных интересов: информационная безопасность, распознавание образов, обработка изображений. Число научных публикаций — 5. gremlin.kop@mail.ru; Приборостроительная, 35, Орел, 302034; р.т.: +7(4862)54-99-33.

Мещеряков Роман Валерьевич – д-р техн. наук, доцент, проректор по научной работе и инновациям, Томский государственный университет систем управления и радиоэлектроники (ТУСУР). Область научных интересов: системный анализ, информационная безопасность, вопросы обработки информации в интеллектуальных системах, информационно-безопасные системы, идентификация диктора, обработка речи, машинное обуче-ние, программно-аппаратные средства защиты информации. Число научных публикаций — 400. mrv@ieee.org; пр. Ленина, 40, Томск, 634050; р.т.: +7(3822)900111, Факс: +7 (3822) 900-111.

Евсютин Олег Олегович — к-т техн. наук, доцент кафедры безопасности информационных систем, Томский государственный университет систем управления и радиоэлектроники (ТУСУР). Область научных интересов: информационная безопасность, обработка цифровых изображений, цифровая стеганография, приложения клеточных автоматов. Число научных публикаций — 50. eoo@keva.tusur.ru; пр. Ленина, 40, Томск, 634050; р.т.: +7(3822)701529, Факс: +7(3822)513262.

Туан Лай Минь — к-т техн. наук, доцент, Академия криптографической техники СШПР Социалистической республики Вьетнам. Область научных интересов: информа-ционная безопасность, защита от несанкционированного доступа. Число научных публикаций — 10. lmtuan.1989@gmail.com; Нгуен Чи Тан, 105, Донг Да, Ханой, 10, Вьетнам; р.т.: +438357975.

Поддержка исследований. Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации в рамках проектной части государственного задания ТУСУР на 2017-2019 гг. (проект № 2.3583.2017/ПЧ)

DOI 10.15622/sp.60.5

A.V. KOZACHOK, S.A. KOPYLOV, R.V. MESHCHERYAKOV,
O.O. EVSUTIN, L.M. TUAN

AN APPROACH TO A ROBUST WATERMARK EXTRACTION FROM IMAGES CONTAINING TEXT

Kozachok A.V., Kopylov S.A., Meshcheryakov R.V., Evsutin O.O., Tuan L.M. **An Approach to a Robust Watermark Extraction from Images Containing Text.**

Abstract. This paper presents an approach to a robust watermark extraction from images containing text. Data extraction based on developed approach to robust watermark embedding into text data, characterizing by conversion invariance of text data into an image format. The comparative analysis of existing approaches of steganographic data embedding into text data is carried out, their advantages and disadvantages are determined. The choice of groups to steganographic data embedding methods based on text formatting is justified. As an embedding algorithm is determined approach based on interline space shifting. The block diagram and the description of the developed algorithm of data embedding into text data are given. An experimental estimation of the embedding capacity and perceptual invisibility of the developed data embedding approach was carried out. An approach to extract embedded information from images containing a robust watermark, based on the existing limitations, has been developed. The Radon transform is chosen as the basic extraction procedure of embedded information, allowing to extract values of the interline spacing. An approach based on Gaussian mixture model separating to isolate the values of the bits was chosen. The limits of the retrieval of embedded data have been experimentally established, and the robustness of the developed embedding approach to the implementation of various transformations has been estimated. The following parameters of robustness developed approach are defined: rotation of an image containing embedded data at any angle; scaling an image with a scaling factor not exceeding 1.5; conversion to any bitmap format; the application of a median filter to an image with a convolution core limit of not more than 9, a Gaussian blur filter with a blurring limit not exceeding 8 and an average filter with a convolution kernel limit of not more than 5.

Keywords: information security, text steganography, digital watermarking, Radon transform, separation of mixture distribution.

Kozachok Alexander Vasilievich — Ph.D., researcher, The Academy of Federal Security Guard Service of the Russian Federation. Research interests: information security, unauthorized access protection, mathematical cryptography, theoretical problems of computer science. The number of publications — 105. alex.totrin@gmail.com; 35, Priborostroitelnaya Street, Orel, 302034, Russia; office phone: +7(486) 254-99-33.

Kopylov Sergey Alexandrovich — researcher, The Academy of Federal Security Guard Service of the Russian Federation. Research interests: information security, protection from unauthorized access, image processing. The number of publications — 5. gremlin.kop@mail.ru; 35, Priborostroitelnaya Street, Orel, 302034, Russia; office phone: +7(4862)54-99-33.

Meshcheryakov Roman Valeryevich — Ph.D., Dr. Sci., professor, vice-rector for research and innovation, Tomsk State University of Control Systems and Radioelectronics (TUSUR). Research interests: speech analysis, speech recognition, medical technology, information security, speaker recognition, speech processing, machine learning, hardware-software data protection solutions. The number of publications — 400. mrv@ieee.org; 40, Lenin-avenue Tomsk, 634050, Russia; office phone: +7(3822)900111, Fax: +7 (3822) 900-111.

Evsutin Oleg Olegovich — Ph.D., associate professor of information system security department, Tomsk State University of Control Systems and Radioelectronics (TUSUR). Research interests: information security, digital images processing, applications of cellular automata theory. The number of publications — 50. eoo@keva.tusur.ru; 40, Lenin-avenue Tomsk, 634050, Russia; office phone: +7(3822)701529, Fax: +7(3822)513262.

Tuan Lai Minh — Ph.D., associate professor, Cryptographic technology Academy. Research interests: information security, unauthorized access protection. The number of publications — 10. lmtuan.1989@gmail.com; 105, Nguen Chi Than, Dong Da, Hanoi, 10, Vietnam; office phone: +438357975.

Acknowledgements. This work is supported by the Ministry of Education and Science of the Russian Federation within the framework of the project part of the state task of TUSUR for 2017-2019 (Project № 2.3583.2017/Pch)

References

1. Weber R.H. Internet of Things – New security and privacy challenges. *Computer Law & Security Review*. 2010. vol. 26. no. 1. pp. 23–30.
2. Chaudhuri A. Internet of things data protection and privacy in the era of the General Data Protection Regulation. *Journal of Data Protection & Privacy*. 2016. vol. 1. no.1. pp. 64–75.
3. JSOC Security flash report 2017 second half year. Available at: solarsecurity.ru/upload/iblock/54f/flash_report_Q2_2017_041.pdf (accessed: 02.05.2018) (In Russ.).
4. Ramapriya B. An Improved Approach of Text Steganography in Application with Rotational Symmetry. *International Journal of Innovative Research in Computer and Communication Engineering*. 2017. vol. 5. no. 7. pp. 12939–12947.
5. Elmahi M.Y., Kosti Su., Sayed M.H. Text Steganography Using Compression and Random Number Generators. *International Journal of Computer Applications Technology and Research*. 2017. vol. 6. no. 6. pp. 259–263.
6. Kingslin S., Kavitha N. Evaluative approach towards text steganographic techniques. *Indian Journal of Science and Technology*. 2015. vol. 8. no. 29. pp. 1–8.
7. Odeh A., Elleithy K., Faezipour M., Abdelfattah E. Highly efficient novel text steganography algorithmss. Systems, Applications and Technology Conference (LISAT). 2015. pp. 1–7.
8. Rathore A.S., Rawat S.S. A Secure Image and Text Steganography Technique. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*. 2015. vol. 3. no. 5. pp. 506–509.
9. Rezanova Z.I., Romanov A.S., Meshcheryakov R.V. [About choice of signs of the text relevant in the autor expert's activity]. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal* 2013. vol. 26. no. 6. pp. 38–52. (In Russ.).
10. Gribunin V.G., Okov I.N., Turincev I.V. *Cifrovaya steganografiya*. [Digital steganography]. M.: SOLON-Press., 2017. 262 p. (In. Russ.).
11. Kozachok A.V., Kopylov S.A., Meshcheryakov R.V., Evsutin O.O. [Robust image hashing survey]. *Kompjuternaja Optika – Computer Optics*. 2017. vol. 41. no. 5. pp. 743–755. (In Russ.).
12. Salomon D. Data privacy and security: encryption and information hiding. *Springer Science & Business Media*. 2003. 469 p.
13. Woo C.-S. Digital image watermarking methods for copyright protection and authentication. Ph.D. Thesis. Queensland University of Technology. 2007. 223 p.
14. Phadikar A. Robust Watermarking Techniques for Color Images. 2009.

15. Evsutin O.O., Shelupanov A.A., Meshcheryakov R.V., Bondarenko D.O. [An algorithm for information embedding into compressed digital images based on replacement procedures with use of optimization]. *Kompjuternaja Optika – Computer Optics.* 2017. vol. 41. no. 3. pp. 412–421. (In Russ.).
16. Vatamanyuk I.V., Ronzhin A.L. [Application of digital images blur estimation methods for audiovisual monitoring]. *Informatsionno-upravliaiushchie sistemy – Information and Control Systems.* 2014. vol. 71. no. 4. pp. 16–23. (In Russ.).
17. Agarwal M. Text steganographic approaches: a comparison. 2013. arXiv preprint arXiv:1302.2718. 16 p.
18. Pamulaparty L., Rao N.M. Text Steganography: Review. *International Journal of Computer Science and Information Technology & Security (IJCSITS)*. 2016. vol. 6. no. 4. pp. 80–83.
19. Jarmolik S.V., Listopad J.N. [Steganographic methods of information protection]. *Informacionnoe obrazovanie – Information education.* 2005. vol. 5. pp. 64–74. (In Russ.).
20. Shirali-Shahreza M., Shirali-Shahreza M.H. Text steganography in SMS. Proceedings of IEEE International Conference on Convergence Information Technology. 2007. pp. 2260–2265.
21. Bondarchuk S.S., Davydova E.M., Kostjuchenko E.J. [Integration of digital characters for information security]. *Doklady Tomskogo Gosudarstvennogo Universiteta Sistem Upravlenija i Radiojelektroniki – Proceedings of Tomsk State University of Control Systems and Radioelectronics*. 2011. vol. 24. no. 2. pp. 228–235. (In Russ.).
22. Govada S.R., Kumar B.S., Devarakonda M., Stephen M.J. Text steganography with multi level shielding. *IJCSI International Journal of Computer Science Issues*. 2012. vol. 9. no. 5. pp. 401–405.
23. Nagarhalli T.P., Bakal J.W., Jain N. A Survey of Hindi Text Steganography. *International Journal of Scientific & Engineering Research*. 2016. vol. 7. no. 3. pp. 55–61.
24. Low S.H., Maxemchuk N.F., Brassil JT., O'Gorman L. Document marking and identification using both line and word shifting. Proceedings of IEEE INFOCOM'95 on Fourteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Bringing Information to People. 1995. vol. 2. pp. 853–860.
25. Saraswathi V., Kingslin M.S. Different Approaches to Text Steganography: A Comparison. *International Journal of Emerging Research in Management & Technology*. 2014. vol. 9359. no. 11. pp. 124–127.
26. Por L.Y., Delina B. Information hiding: A new approach in text steganography. Proceedings of WSEAS International Conference on Mathematics and Computers in Science and Engineering. 2008. vol. 7. pp. 689–695.
27. Villán R. et al. Tamper-proofing of electronic and printed text documents via robust hashing and data-hiding. *Security, Steganography, and Watermarking of Multimedia Contents*. 2007. pp. 6505–6517.
28. Kumar K.A., Pabboju S., Desai N.M.S. Advance text steganography algorithms: an overview. *International Journal of Research and Applications*. 2014. vol. 1. no. 1. pp. 31–35.
29. Bhaya W., Rahma A.M., Al-nasrawi D. Text steganography based on font type in MS-Word documents. *Journal of Computer Science*. 2013. vol. 9. no. 7. pp. 898–904.
30. Rahma A.M.S., Bhaya W.S., Al-Nasrawi D.A. Text steganography based on unicode of characters in multilingual. *International Journal of Engineering Research and Applications*. 2013. vol. 3. no. 4. pp. 1153–1165.
31. Word 2007: Rich Text Format (RTF) Specification, version 1.9.1. Available at: www.microsoft.com/en-us/download/details.aspx?id=10725 (accessed: 02.05.2018).
32. 171/SC 2: ISO 32000–1: 2008 Document Management-Portable Document Format-Part 1. ISO. TC. 2008. 747 p.

33. 171/SC 2: ISO 32000–2: 2017 Document Management-Portable Document Format-Part 2. ISO. TC. 2017. 371 p.
34. Toft P.A., Sørensen J.A. The Radon transform-theory and implementation. Ph.D. Thesis. Technical University of Denmark. 1996. 321 p.
35. Volkov V.A. *Rjady Fur'e. Integral'noe preobrazovanie Fur'e i Radona: uchebno-metodicheskoe posobie*. [Fourier series. Integral Fourier and Radon transform: a teaching aid]. Ekaterinburg: Ural University Publ., 2014. 32 p. (In. Russ.)
36. Bahaghighat M.K., Mohammadi J. Novel approach for baseline detection and Text line segmentation. *International Journal of Computer Applications*. 2012. vol. 51. no. 2. pp. 9–16.
37. GOST R IEC/TR 61948-2–2008 [Nuclear medicine instrumentation. Routine tests. Part 2. Scintillation cameras and single photon emission computed tomography imaging] M.: Gosstandart Rossii. 2009. 8 p. (In. Russ.)
38. Reynolds D. Gaussian mixture models. Encyclopedia of biometrics. 2015. pp. 827–832.
39. Ajvazjan S.A., Buhshtarev V.M., Enjukov E.S. *Prikladnaja statistika. Klassifikacija i snizhenie razmernosti*. [Applied statistics. Classification and reduction of dimension]. M.: Finance and statistics. 1989. 607 p. (In. Russ.)