DOI: 10.15514/ISPRAS-2022-34(4)-11

Экспериментальная оценка алгоритма маркирования текстовых документов на основе изменения интервалов между словами

[1] *А.В. Козачок, ORCID: 0000-0002-6501-2008 <a.kozachok@academ.msk.rsnet.ru>*
[2] *В.И. Козачок, ORCID: 0000-0001-5384-2269 <v.kozachok@academ.msk.rsnet.ru>*
[2] *С.А. Копылов, ORCID: 0000-0003-2841-5243 <gremlin.kop@mail.ru>*
[2] *П.Н. Горбачев, ORCID: 0000-0002-4511-0348 <png@academ.msk.rsnet.ru>*
[1] *Ю.В. Маркин, ORCID: 0000-0003-1145-5118 <ustas@ispras.ru>*
[1] *Д.О. Обыденков, ORCID: 0000-0002-9296-6333 <obydenkov@ispras.ru>*
[1] *Институт системного программирования им. В.П. Иванникова РАН, 109004, Россия, г. Москва, ул. А. Солженицына, д. 25*
[2] *Академия Федеральной службы охраны Российской Федерации, 302015, Россия, г. Орёл, ул. Приборостроительная, д. 35*

Аннотация. В статье представлены результаты экспериментальной оценки параметров алгоритма маркирования электронных документов, основанного на изменении интервалов между словами. Разработанный алгоритм маркирования предназначен для повышения защищенности электронных документов, содержащих текстовую информацию, от утечки по каналам, обусловленным печатью, сканированием или фотографированием и последующей отправкой сформированного изображения. В качестве анализируемых параметров алгоритма выступают такие характеристики как емкость встраивания, невидимость, необнаруживаемость, извлекаемость и робастность. В ходе оценки емкости встраивания разработанного алгоритма приведены аналитические выражения, позволяющие рассчитать величину предельно достижимой емкости встраивания. Полученные количественные оценки и проведенные эксперименты позволили обосновать выбор допустимых значений встраиваемого маркера. Для определения невидимости встраиваемой информации в исходный документ осуществлена оценка невидимости и необнаруживаемости встроенного маркера. В ходе проведения экспертной оценки обоснована невидимость разработанного алгоритма к визуальному анализу, а также отсутствие значительных статистических отклонений в распределении анализируемых параметров в процессе оценки стойкости разработанного алгоритма маркирования к потенциально наилучшему методу стеганографического анализа. Получены количественные значения извлекаемости разработанного алгоритма маркирования посредством оценки точности извлечения. Проведенный анализ показал высокие значения точности извлечения маркера из отсканированных изображений, позволяющие гарантированно извлекать встроенные данные, а также направления совершенствования извлекаемости маркера из сфотографированных изображений. В процессе оценки устойчивости разработанного алгоритма маркирования к осуществлению преобразований и внесению искажений определены основные параметры робастности разработанного алгоритма маркирования к процессам печати, сканирования и фотографирования. Сформулированы выводы о возможности применения разработанного алгоритма маркирования и направления дальнейших исследований.

Ключевые слова: защита от утечки информации; маркирование; распознавание образов; обработка изображений; стеганографический анализ

Для цитирования: Козачок А.В., Козачок В.И., Копылов С.А., Горбачев П.Н., Маркин Ю.В., Обыденков Д.О. Экспериментальная оценка алгоритма маркирования текстовых документов на основе изменении интервала между словами. Труды ИСП РАН, том 34, вып. 4, 2022 г., стр. 153-172. 10.15514/ISPRAS-2022-34(4)-11

Experimental evaluation of the text documents marking algorithm based on interword distances shifting

[1] *A.V. Kozachok, ORCID: 0000-0002-6501-2008<a.kozachok@academ.msk.rsnet.ru>*
[2] *V.I. Kozachok, ORCID: 0000-0001-5384-2269 <v.kozachok@academ.msk.rsnet.ru>*
[2] *S.A. Kopylov, ORCID: 0000-0003-2841-5243 <gremlin.kop@mail.ru>*
[2] *P.N. Gorbachev, ORCID: 0000-0002-4511-0348 <png@academ.msk.rsnet.ru>*
[1] *Y.V. Markin, ORCID: 0000-0003-1145-5118 <ustas@ispras.ru>*
[1] *D.O. Obydenkov, ORCID: 0000-0002-9296-6333 <obydenkov@ispras.ru>*
[1] *Ivannikov Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia*
[2] *Academy of Federal Guard Service, 35, Priborostroitel'naya st., Orel, 302015, Russia*

Abstract. The article presents the experimental parameter evaluation results of the electronic documents marking algorithm, based on interword distances shifting. The developed marking algorithm is designed to increase the security of electronic documents containing textual information from leakage through channels caused by printing, scanning or photographing, followed by sending the generated image. The algorithm analyzed parameters are such characteristics as embedding capacity, invisibility, undetectability, extractability and robustness. In the course of embedding capacity estimation of the developed algorithm, analytical expressions are given that make it possible to calculate the maximum achievable embedding capacity value. The obtained quantitative estimates and the experiments carried out made it possible to substantiate the admissible values choice of the embedded marker. To determine the embedded information invisibility in the source document, an invisibility and undetectability assessment of the embedded marker was carried out. During the expert evaluation, the developed algorithm invisibility to visual analysis was substantiated, as well as the absence of significant statistical deviations in the distribution of the analyzed parameters in the process of assessing the resistance of the developed marking algorithm to the potentially best steganographic analysis method. The quantitative extractability of the developed marking algorithm was carried out by assessing the extraction accuracy. The analysis performed showed accuracy high values of marker extraction from scanned images, which makes it possible to reliably extract embedded data, as well as determine directions for improving the extraction accuracy from photographed images. In the assessing process the stability of the developed marking algorithm to the transformations implementation and distortions introduction, the main robustness parameters of the developed marking algorithm to the printing, scanning and photographing processes are determined. Conclusions are formulated on the using possibility the developed marking algorithm and directions for further researches are identified.

Keywords: information leakage protection; marking; pattern recognition; image processing; steganographic analysis

For citation: Kozachok A.V., Kozachok V.I., Kopylov S.A., Gorbachev P.N., Markin Y.V., Obydenkov D.O. Experimental evaluation of the text documents marking algorithm based on interword distances shifting. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 4, 2022. pp. 153-172 (in Russian). DOI: 10.15514/ISPRAS-2022-34(4)-11

1. Введение

В последнее десятилетие задача по обеспечению безопасности конфиденциальной информации и данных стала одной из наиболее актуальных проблем в области информационной безопасности. С ростом количества информации и данных, обрабатываемых и циркулирующих в информационно-телекоммуникационных сетях, возросло и количество инцидентов информационной безопасности. Основной вектор нарушений в области информационной безопасности приходится на действия внешних нарушителей. Так, в 2021 году злоумышленниками осуществлено 2418 компьютерных атак на информационные ресурсы частных лиц (349 случаев) и информационно-коммуникационные сети госучреждений (322 случая) [1]. Несмотря на значительное

превосходство по количеству (1729 против 2418 случаев), утечки конфиденциальной информации, реализованные внутренними нарушителями, привели к компрометации 8,42 млрд. записей конфиденциальной информации и данных [2]. Высокие значения скомпрометированной информации позволяют отнести задачу по обеспечению защиты информации от утечки к наиболее актуальным направлениям исследований.

В результате анализа каналов утечки установлено, что наибольшее число случаев нарушений информационной безопасности связано с отправкой конфиденциальной информации по сети и электронной почте (89,6 % случаев от общего числа). При этом основным типом данных подвергшимся компрометации являются изображений печатных текстовых документов, полученных посредством сканирования или фотографирования. Наличие указанного канала утечки обусловлено отсутствием механизмов маркирования и обнаружения печатных документов в современных DLP-системах [3–6].

В ходе решения задачи по повышению защищенности конфиденциальной информации от утечки по каналу, обусловленному сканированием или фотографированием напечатанных документов с последующей отправкой по сети, разработан алгоритм маркирования электронных документов, основанный на внедрении маркера за счет горизонтального сдвига слов (символов) [7–8].

Алгоритм маркирования электронных текстовых документов, выводимых на печать, основан на применении технологии выделения строк текста из изображения, с последующим выделением слов (символов) внутри строк, формирования из последних областей встраивания и встраивания в сформированные области метки за счет изменения интервалов между словами в каждой строке. Встраивание информации осуществляется за счет внедрения в строку, содержащую более трех слов, пробела установленной длины (удлиненного пробела) с компенсацией (увеличением или уменьшением) остальных величин интервалов между словами в строке.

Практическая реализация прототипа алгоритма маркирования позволила оценить применимость разработанного алгоритма к маркированию различных типов данных. Так разработанный прототип реализует маркирование не только электронных текстовых документов формата: .doc, .docx, .rtf, .pdf, odt, .fodt но и электронных изображений, содержащих текстовые области, вне зависимости от используемого формата и алгоритма сжатия. Реализованный прототип позволяет провести экспериментальную оценку разработанного алгоритма маркирования и сделать вывод о возможности повышения защищенности электронных документов от утечки по каналу, обусловленному сканированием или фотографированием напечатанных на бумаге документов.

2. Экспериментальная оценка параметров алгоритма маркирования электронных документов

Экспериментальная оценка разработанного алгоритма осуществлялась на основе количественной и качественной оценки основных параметров, характеризующих встраиваемый маркер (алгоритм маркирования). К таким параметрам относятся [9–11]:

- емкость встраивания (полезная нагрузка);
- невидимость (перцептивная прозрачность);
- необнаруживаемость (сложность обнаружения);
- извлекаемость;
- робастность.

2.1 Емкость встраивания (полезная нагрузка)

Емкость встраивания (полезная нагрузка) – количество информации, которое может быть встроено (внедрено) в исходный документ [12, 13]. Основной величиной, характеризующей максимально возможное количество информации, которое может быть встроено в исходный документ, является предельно достижимая емкость встраивания. Предельно достижимая емкость встраивания η разработанного алгоритма маркирования рассчитывается по формуле:

$$\eta = \sum_{l=1}^{N} |DS[l]|, \quad (1)$$

где N – количество строк электронного документа, l – строка текста ($l = 1,2,\ldots N$), DS – блок, состоящий из двух последовательно следующих пробелов (интервалов между словами).

Количество строк текстового документа полностью заполненного текстом N может быть рассчитано посредством выражения [14]:

$$N = \left\lfloor \frac{H - (m_t - m_b}{\gamma + \beta \cdot \gamma} \right\rfloor, \quad (2)$$

где H – высота текстового документа, m_t, m_b – размер верхнего и нижнего поля текстового документа соответственно, γ – размер кегля шрифта, β – величина множителя межстрочного интервала.

Для электронных документов, оформленных в соответствии с требования ГОСТ Р 7.0.97–2016 [15], значения высоты текстового документа H, размеров верхнего и нижнего поля текстового документа m_t, m_b представляют фиксированные величины со следующими значениями:

- высота текстового документа: формат А4 – 297 мм., А5 – 210 мм.;
- поля листа документа: верхнее – 20 мм., нижнее – 20 мм.;
- размер кегля шрифта: 12, 13 и 14 пт;
- величина множителя межстрочного интервала: 1–1,5.

Исходя из представленных значений, рассчитаны предельные значения количества строк страницы текстового документа полностью заполненной текстом:

- формат А4: от 52 (межстрочный интервал 1, кегль шрифта 12 пт) до 30 строк (межстрочный интервал 1,5, кегль шрифта 14 пт).;
- формат А5: от 34 (межстрочный интервал 1, кегль шрифта 12 пт) до 19 строк (межстрочный интервал 1,5, кегль шрифта 14 пт).

Помимо количества строк, приходящихся на страницу электронного документа, полностью заполненного текстом, осуществлен расчет среднего числа пробелов в строке. В результате проведенного анализа получены следующие значения:

- шрифт с кеглем 12 пт – 9 пробелов (А4), 6 пробелов (А5);
- шрифт с кеглем 13 пт – 8 пробелов (А4), 5 пробелов (А5);
- шрифт с кеглем 14 пт – 7 пробелов (А4), 4 пробела (А5).

Полученные оценки количества строк страницы текстового документа полностью заполненной текстом N, а также среднего числа пробелов в строке позволяют рассчитать значения предельно достижимой емкости встраивания η в зависимости от параметров оформления текстового документа. Полученные значения представлены в таблицах 1 и 2.

Табл. 1. Величина предельно достижимой емкости встраивания страницы текста формата А4
Table 1. The maximum achievable capacity value for embedding of an A4 text page

Кегль шрифта γ (пт)	**Величина межстрочного интервала β (множитель)**	**Количество строк N**	**Предельно достижимая емкость встраивания η**
12	1	52	208
	1,25	42	168
	1,5	35	140
13	1	48	192
	1,25	38	152
	1,5	32	128
14	1	45	135
	1,25	36	108
	1,5	30	90

Табл. 2. Величина предельно достижимой емкости встраивания страницы текста формата А5
Table 2. The maximum achievable capacity value for embedding of an A5 text page

Кегль шрифта γ (пт)	**Величина межстрочного интервала β (множитель)**	**Количество строк N**	**Предельно достижимая емкость встраивания η**
12	1	34	102
	1,25	27	81
	1,5	23	69
13	1	32	64
	1,25	25	50
	1,5	21	42
14	1	29	58
	1,25	23	46
	1,5	19	38

Анализ полученных результатов позволяет сделать вывод о том, что предельно достижимая емкость встраивания разработанного алгоритма маркирования составляет 208 бит для листа формата А4 и 102 для формата А5. При этом указанные значения достигаются посредством полного заполнения текстом страницы электронного документа. Полученные значения емкости встраивания разработанного алгоритма позволяют встраивать маркер размером 64 бита для документов формата А4, и 32 бита – формата А5. Для обоснования оптимального размера маркера необходимо оценить невидимость и необнаруживаемость встроенных данных (маркера) разработанного алгоритма маркирования.

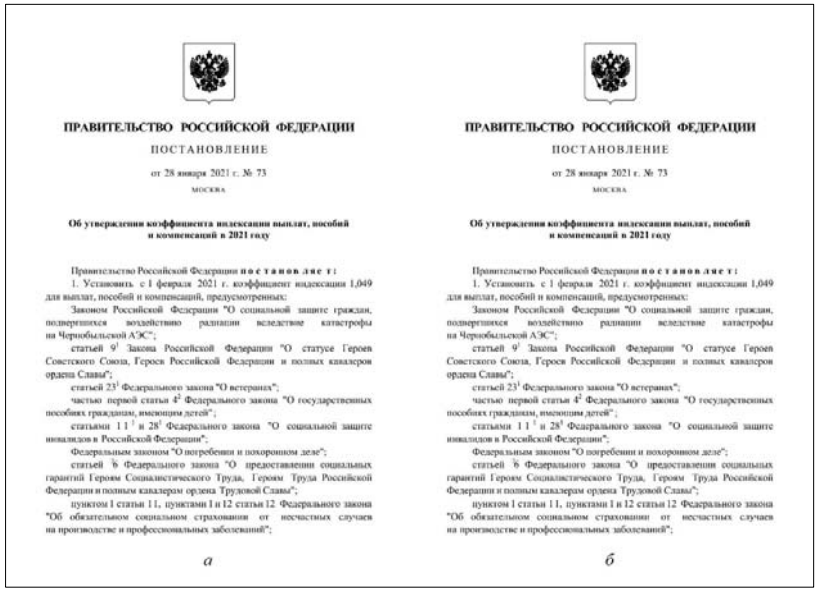
2.2 Невидимость (перцептивная прозрачность)

Невидимость (перцепционная прозрачность) встроенной информации (маркера) – качественная характеристика, отражающая степень искажения контейнера встраиваемыми данными. Данная характеристика основана на перцептивном восприятии человека и может быть оценена посредством проведения визуального анализа наблюдателем [16, 17].

В ходе оценки невидимости разработанного алгоритма маркирования проведены три группы исследований, направленных на оценку зависимости невидимости встроенного маркера от величины удлиненного пробела посредством визуального анализа. Визуальный анализ проводился посредством экспертного зрительного анализа подписанных электронных

документов, содержащих текст, представленных как в напечатанном, так и в электронном виде (в формате отсканированных изображений).

Внедрение маркера осуществлялось в электронные изображения, содержащие помимо текстовой информации различные графические объекты, таблицы, а также элементы рукописных подписей и печатей. Пример оригинального и соответствующего ему маркированного изображения представлен на рис. 1.


ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

ПОСТАНОВЛЕНИЕ

от 28 января 2021 г. № 73

МОСКВА

Об утверждении коэффициента индексации выплат, пособий и компенсаций в 2021 году

Правительство Российской Федерации **п о с т а н о в л я е т :**

1. Установить с 1 февраля 2021 г. коэффициент индексации 1,049 для выплат, пособий и компенсаций, предусмотренных:

Законом Российской Федерации "О социальной защите граждан, подвергшихся воздействию радиации вследствие катастрофы на Чернобыльской АЭС";

статьей 9^1 Закона Российской Федерации "О статусе Героев Советского Союза, Героев Российской Федерации и полных кавалеров ордена Славы";

статьей 23^1 Федерального закона "О ветеранах";

частью первой статьи 4^2 Федерального закона "О государственных пособиях гражданам, имеющим детей";

статьями 11^1 и 28^1 Федерального закона "О социальной защите инвалидов в Российской Федерации";

Федеральным законом "О погребении и похоронном деле";

статьей 6 Федерального закона "О предоставлении социальных гарантий Героям Социалистического Труда, Героям Труда Российской Федерации и полным кавалерам ордена Трудовой Славы";

пунктом 1 статьи 11, пунктами 1 и 12 статьи 12 Федерального закона "Об обязательном социальном страховании от несчастных случаев на производстве и профессиональных заболеваний";

а

ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

ПОСТАНОВЛЕНИЕ

от 28 января 2021 г. № 73

МОСКВА

Об утверждении коэффициента индексации выплат, пособий и компенсаций в 2021 году

Правительство Российской Федерации **п о с т а н о в л я е т :**

1. Установить с 1 февраля 2021 г. коэффициент индексации 1,049 для выплат, пособий и компенсаций, предусмотренных:

Законом Российской Федерации "О социальной защите граждан, подвергшихся воздействию радиации вследствие катастрофы на Чернобыльской АЭС";

статьей 9^1 Закона Российской Федерации "О статусе Героев Советского Союза, Героев Российской Федерации и полных кавалеров ордена Славы";

статьей 23^1 Федерального закона "О ветеранах";

частью первой статьи 4^2 Федерального закона "О государственных пособиях гражданам, имеющим детей";

статьями 11^1 и 28^1 Федерального закона "О социальной защите инвалидов в Российской Федерации";

Федеральным законом "О погребении и похоронном деле";

статьей 6 Федерального закона "О предоставлении социальных гарантий Героям Социалистического Труда, Героям Труда Российской Федерации и полным кавалерам ордена Трудовой Славы";

пунктом 1 статьи 11, пунктами 1 и 12 статьи 12 Федерального закона "Об обязательном социальном страховании от несчастных случаев на производстве и профессиональных заболеваний";

б

Рис. 1. Изображение: а) не содержащее встроенные данные; б) содержащее маркер
Fig. 1. Image: a) not containing embedded data; b) containing a marker

Первая группа исследований направлена на определение факта наличия встроенной информации, содержащейся в анализируемых подписанных изображениях (изображениях, содержащих встроенный маркер). В данной группе исследований аналитики (эксперты по стеганографическому анализу) не обладали информацией о содержании в анализируемых изображениях встроенных данных. В качестве анализируемых изображений выступали изображения, в которых осуществлено внедрение маркера посредством изменения величины удлиненного пробела разработанным алгоритмом маркирования. В результате анализа подписанных изображений не определен факт наличия встроенных данных.

При проведении второй группы исследований аналитики обладали информацией о факте наличия в анализируемых изображениях встроенных данных. При этом алгоритм внедрения информации оставался неизвестным. В результате проведенного визуального анализа не были обнаружены аномалии в структуре текста подписанных изображений, что позволило сделать вывод о невидимости встроенные данных к данному виду анализа.

В ходе проведения третьей группы исследований аналитики обладали информацией об используемом алгоритме внедрения информации, при этом место встраивания (строки текста) для аналитиков оставалось неизвестным. В результате визуального анализа аналитиками обнаружена каждая третья строка текста, содержащая удлиненный пробел. В остальных случаях позиции удлиненного пробела определены ошибочно или не определены совсем.

Результаты визуального анализа позволяют сделать вывод о стойкости разработанного подхода маркирования текстовых документов к обнаружению посредством визуального анализа встроенного маркера. Для определения статистических отклонений в сформированных в результате маркирования изображениях необходимо оценить необнаруживаемость (сложность обнаружения) разработанного алгоритма маркирования.

2.3 Необнаруживаемость (сложность обнаружения)

Под необнаруживаемостью (сложностью обнаружения) понимается количественная характеристика, отражающая степень искажения статистических характеристик контейнера, не связанных с перцептивным восприятием человека. Необнаруживаемость может быть оценена как стойкость разработанного алгоритма маркирования к потенциально наилучшему методу стеганографического анализа [18, 19].

Стеганографический анализ (стегоанализ, стеганоанализ) – наука об обнаружении факта присутствия (наличия) скрываемой информации в анализируемых контейнерах (объектах) [20, 21]. В качестве контейнера (объекта) могут выступать данные мультимедиа: изображения, аудио и видео данные. Помимо обнаружения факта сокрытия данных внутри анализируемых объектов дополнительной целью стегоанализа может выступать извлечение встроенной информации. Исходя из особенностей разработанного алгоритма маркирования, в качестве анализируемого контейнера выступают изображения, полученные посредством сканирования или фотографирования напечатанных на бумаге текстовых документов. Контейнеры отличные от указанных в работе не рассматриваются.

Существующие методы стеганографического анализа изображений можно разделить на специальные и общие (слепой стегоанализ) [22]. Обнаружение встроенной информации специальными методами осуществляется только при наличии информации (в том числе априорной) об используемом алгоритме стеганографического внедрения информации. К специальным методам относятся сигнатурные и схемные. Достоинством методов данной группы является возможность извлечения встроенной информации из подписанного изображения. При этом, методы данной группы малоэффективны против новых или неизвестных алгоритмов стеганографического внедрения информации. Сигнатурные методы стегоанализа основаны на поиске в анализируемом изображении сигнатуры (шаблона), характеризующего конкретный алгоритм или программное средство, осуществляющее стеганографическое внедрение информации. Методы сигнатурного стегоанализа позволяют однозначно определить факт встраивания и идентифицировать используемый стеганографический алгоритм (программное средство). При этом количество программных средств внедрения информации, имеющих собственные сигнатуры, исчисляется несколькими десятками, что не позволяет использовать данный подход для обнаружения новых (неизвестных) алгоритмов (программных средств).

Анализ существующих подходов к стеганографическому анализу позволяет обосновать выбор в качестве потенциально наилучших методов стеганографического анализа следующие подходы: визуальный стегоанализ и метод статистического стегоанализа в пространственной области, основанный на анализе гистограмм, построенных по частотам пробелов изображения. Основываясь на полученных результатах оценки невидимости (перцепционной прозрачности), сделан вывод о стойкости разработанного алгоритма маркирования к проведению визуального стегоанализа.

В ходе экспериментальной оценки стойкости разработанного подхода к методу статистического стегоанализа в пространственной области, основанного на анализе гистограмм, построенных по частотам пробелов изображения (потенциально наилучший метод стегоанализа), осуществлено извлечение информации как из неподписанных изображений (не содержащих встроенные данные), так и из изображений, содержащих встроенный маркер.

На рис. 2 представлен пример маркированных текстовых электронных документов (где 2*а* – сканированная версия документа, 2*б* – конвертированное изображение), которые совместно с оригинальными документами подвергались статистическому стегоанализу в пространственной области, основанному на анализе гистограмм.

Рис. 2. Примеры немаркированных документов
Fig. 2. Examples of unmarked documents

Результаты извлечения величин пробелов (интервалов между словами) из оригинальных изображений и маркированных копий (рис. 2), представлены в таблице 3.

Табл. 3. Результат извлечения интервалов между словами
Table 3. Extracting intervals between words result

Изображение	Извлеченные значения интервалов между словами (в пикселях)
Рис. 2*а* (оригинальный)	27,2; 16,3; 16,3; 16,3; 26,8; 18,1; 18,1; 18,1; 18,1; 26,8; 20,1; 20,1; 26,5; 20,1; 20,1; 16,2; 26,8; 21,8; 21,8; 28,8 21,8; 21,8; 28,8; 21,8; 21,8; 21,8; 79,8; 62,1; 62,1; 62,1; 79,8; 62,1; 29,9; 29,9; 29,9; 40,1; 29,9; 40,1; 41,3; 31,9; 31,9; 31,9; 51,34; 67,8; 51,3; 51,3; 51,3; 51,3; 51,3; 67,8; 67,8...
Рис. 2*а* (маркированный)	19; 18; 18; 21; 19; 23; 22; 18; 21; 23; 20; 20; 21; 22; 24; 21; 22; 23; 24; 22; 23; 24; 24; 24; 24; 22; 68; 68; 69; 65; 68; 70; 34; 34; 33; 34; 33; 32; 34; 36; 34; 33; 57; 55; 58; 55; 57; 55; 57; 57; 55; 57; 56; 56; 56; 54; 74; 74; 75; 72; 74; 74; 74; 74; 79; 52; 52; 52; 48; 51; 53; 39; 39; 35; ...
Рис. 2*б* (оригинальный)	47,5; 26,5; 26,8; 20,5; 20,5; 20,5; 20,5; 21,7; 29,2; 21,7; 21,7; 21,7; 19,1; 19,1; 26,8; 19,1; 19,1; 26,8; 20,4; 20,4; 27,8; 20,4; 19,8; 28,5; 19,8; 19,8; 38,9; 30,3; 30,3; 30,3; 30,3; 24,5; 33,5; 24,5; 24,5; 24,5; 33,5; 23,4; 23,4; 32,5; 23,4; 32,5; 23,4; 23,4; 67,9; 82,5; 67,9; 67,9; 67,9; 18,6; 18,6; ...
Рис. 2*б* (маркированный)	36; 38; 21; 24; 19; 22; 23; 24; 20; 24; 24; 24; 22; 23; 22; 23; 19; 21; 23; 22; 22; 22; 19; 23; 23; 23; 33; 31; 29; 34; 33; 30; 26; 26; 29; 27; 27; 27; 27; 27; 24; 26; 27; 24; 69; 70; 71; 73; 71; 21; 23; 22; 22; 22; 22; 29; 28; 32; 33; 31; 30; 61; 60; 61; 61; 58; 60; 58; 59; 57; 28; 27; 28; 24; ...

Анализ статистических значений полученных величин интервалов между словами позволяет сделать вывод об отсутствии статистической зависимости в распределении интервалов между словами как в оригинальных (немаркированных), так и в маркированных изображениях. На основе полученных массивов величин интервалов между словами осуществлено построение гистограмм (функции плотности распределения вероятностей). Результат построения гистограмм распределения интервалов между словами оригинального и маркированного изображения 2*а* и 2*б* представлены на рисунках 3 и 4 соответственно.

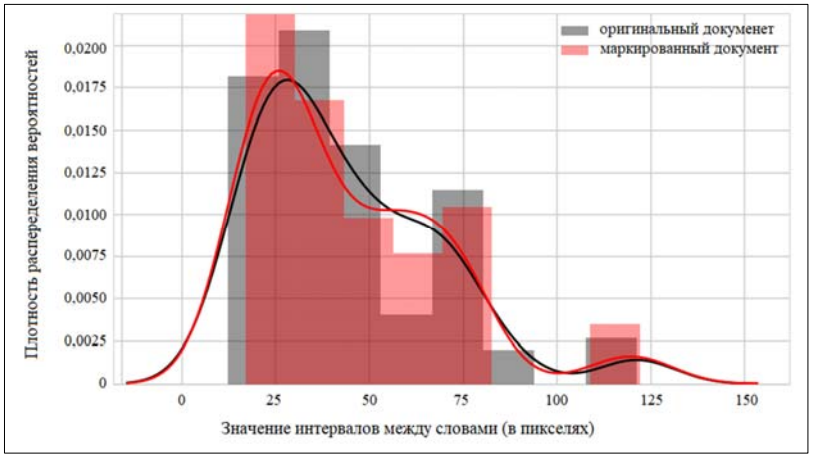


Рис. 3. Гистограмма распределения значений интервалов между словами оригинального и маркированного изображения 2а

Fig.3. Interval values distribution histogram between the words of the original and marked image 2a

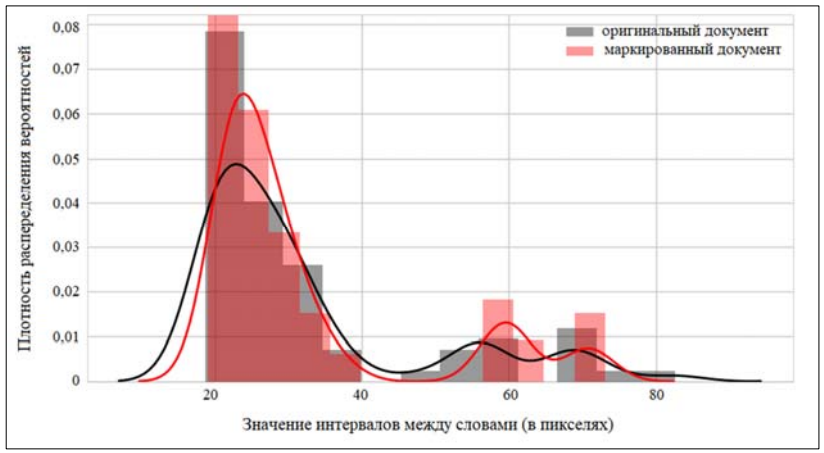


Рис. 4. Гистограмма распределения значений интервалов между словами оригинального и маркированного изображения 2б

Fig.4. Interval values distribution histogram between the words of the original and marked image 2b

Сравнительный анализ гистограмм распределения величин интервалов между словами позволяет сделать вывод и подобие (сходстве) гистограммы маркированного документа оригиналу. Стоит отметить, что имеются незначительные аномалии в сформированных гистограммах, которые обусловлены процессами конвертации, сканирования или

фотографирования изображения. Помимо аномалий указанные преобразования вносят изменения в величины извлеченных значений интервалов между словами в пикселях ввиду изменения разрешения изображения.

Результаты стойкости разработанного алгоритма маркирования электронных документов к потенциально наилучшему методу стегоанализа позволяют сделать вывод об стойкости разработанного алгоритма к указанному методу стегоанализа. Полученные значения невидимости и необнаруживаемости разработанного алгоритма маркирования позволяют перейти к оценке извлекаемости встроенной информации.

2.4 Извлекаемость

Извлекаемость – способность правильного извлечения встроенных данных из контейнера [23]. В ходе оценки извлекаемости встроенных данных из изображений было осуществлено встраивание маркера в текстовые электронные документы (формат .doc, .docx, .pdf, .rtf, .odt, .fodt) и извлечение встроенного маркера (формат .png) со следующими параметрами:

- кегль шрифта: 12 пт, 13 пт и 14 пт;
- межстрочный интервал: 1; 1,25 и 1,5.

Для экспериментальной оценки зависимости точности извлечения встроенных данных от параметров оформления электронных документов, осуществлено: встраивание маркера в подготовленные электронные документы, содержащие текст, преобразование (конвертация) электронных документов из формата .doc, .docx, .pdf, .rtf, .odt, .fodt в формат изображения PNG и извлечение встроенной информации. При преобразовании электронного документа в формат изображения использовано значение растеризации (разрешение изображения), равное 300 точек на дюйм.

Количественно извлекаемость может быть выражена посредством точности извлечения встроенных данных, которая описывается посредством методов оценки классификации, используемых в задачах теории распознавания образов [133–135]. Для количественной оценки точности извлечения рассмотрены следующие метрики точности классификатора: Accuracy и F-мера:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN};\ F - \text{мера} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN},$$

где TP – истинно-положительный, TN – истинно-отрицательный, FP – ложно-положительный и FN – ложно-отрицательный результат.

В процессе экспериментальной оценки было извлечено более 15 000 бит (более 350 страниц текстовой информации), что позволяет утверждать о том, что доверительный вероятность равна 0,95 при точности 0,01. Результаты зависимости точности извлечения от параметров оформления текстовых документов (размер кегля шрифта, величина множителя межстрочного интервала) представлены в табл. 4.

Табл. 4. Точность извлечения данных из изображений в зависимости от параметров оформления текстовых документов

Table 4. Data extraction accuracy from images depending on the design parameters of text documents

Кегль шрифта γ (пт)	**Величина межстрочного интервала β (множитель)**	**Точность излечения F-мера**	**Ложные срабатывания**	**Пропуск цели**
12	1	0,963	0,018	0,019
	1,25	0,961	0,015	0,024
	1,5	0,964	0,020	0,016
13	1	0,966	0,017	0,017
	1,25	0,965	0,015	0,020

	1,5	0,970	0,020	0,010
14	1	0,978	0,014	0,018
	1,25	0,972	0,014	0,014
	1,5	0,980	0,012	0,008

Полученные значения точности извлечения встроенного маркера из изображений позволяют сделать вывод об отсутствии зависимости результата извлечения встроенной информации от параметров электронного документа (размера кегля шрифта, величины межстрочного интервала). Результат извлечения встроенной информации характеризуется одиночными ошибками извлечения первого и второго рода и позволяет сделать вывод о возможности извлечения встроенной информации разработанным алгоритмом из подписанных изображений со значение точности извлечения более 0,95.

В ходе оценки зависимости точности извлечения от разрешения сформированного изображения проведены три группы исследований. В рамках первой группы осуществлена количественная оценка точности извлечения в зависимости от разрешения изображения, сформированного посредством конвертации электронного документа, содержащего встроенный маркер, в изображение. В рамках второй группы оценена зависимость точности извлечения от разрешения изображения, сформированного посредством печати и сканирования электронного документа, содержащего встроенные данные. Третья группа исследований позволила оценить зависимость точности извлечения от разрешения изображения, сформированного посредством печати и фотографирования электронного документа, содержащего встроенные данные.

В процессе количественной оценки точности использованы следующие значения разрешения изображений: 150, 200, 300, 400, 500, 600 и 1200 точек на дюйм. Результаты точности извлечения встроенной информации из изображений, содержащих текст с размером шрифта 12 пт, величиной межстрочного интервала 1, представлены в табл. 5.

Табл. 5. Результаты извлечения информации из конвертированного изображения, содержащего встроенный маркер

Table 5. Results of extracting information from a converted image containing an embedded marker

Разрешение изображения	**Показатель истинно-положительных значений**	**Показатель истинно-отрицательных значений**	**Accuracy**	**F-мера**
150	0,97	0,95	0,96	0,96
200	0,97	0,97	0,97	0,97
300	1	0,97	0,97	0,97
400	1	0,97	0,98	0,98
500	1	0,98	0,99	0,99
600	1	1	1	1
1200	1	0,98	0,99	0,99

Анализ полученных результатов точности извлечения позволяет сделать вывод о наличие незначительной зависимости значений точности извлечения от используемых параметров электронного документа, содержащего текст, и наличие зависимости от разрешения сформированного изображения. При этом стоит отметить тот факт, что показатели точности извлечения для изображений с разрешением в 150 и 200 точек на дюйм, характеризуются наличием ошибок первого и второго рода. В то время как изображения с расширением в 300 точек на дюйм и выше имеют незначительный процент ошибок первого рода, что соответствует наличию одиночных ошибок, которые могут быть устранены посредством применения помехоустойчивых кодов в процессе маркирования.

В ходе второй группы экспериментов осуществлена оценка зависимости точности извлечения встроенных данных от разрешения отсканированного изображения и параметров оформления документов. Для этого осуществлено встраивание маркера в электронные документы, содержащие текст, печать электронных документов, сканирование напечатанных документов и излечение встроенной информации. В качестве используемой метрики оценки точности извлечения информации использована F-мера. В процессе сканирования напечатанных электронных документов использовались следующие значения: 150, 200, 300, 400, 500 и 600 точек на дюйм.

Результаты точности извлечения встроенной информации из электронных изображений, полученных посредством печати и сканирования электронных документов, оформленных в соответствии с требованиями ГОСТ 7.0.97–2016, представлены на рис. 5.

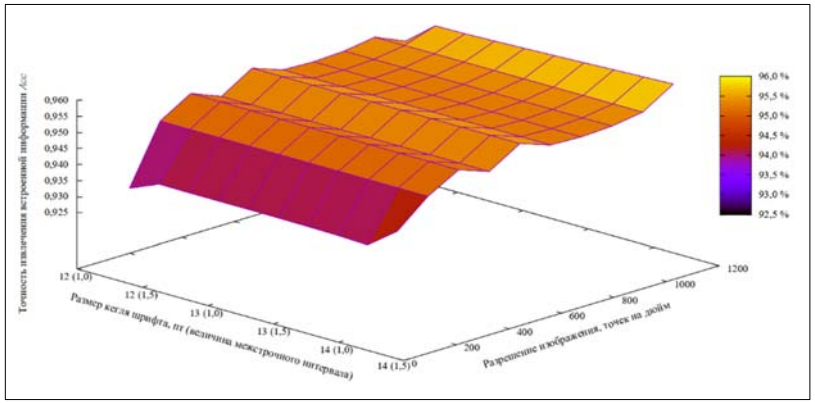


Рис. 5. Зависимость точности извлечения от параметров отсканированного изображения и оформления текстового документа, содержащего встроенный маркер

Fig.5. Extraction accuracy dependency on the parameters of the scanned image and text document design containing an embedded marker

Полученные значения точности извлечения встроенных данных из отсканированных изображений позволяет отнести разработанный алгоритм маркирования к алгоритмам с высокой точностью извлечения (близкой или равной 100%) ввиду наличия ошибок извлечения, которые могут быть исправлены в случае применения методов помехоустойчивого кодирования. Указанные значения точности извлечения достигаются для электронных документов, содержащих текст и оформленных в соответствии требований ГОСТ 7.0.97–2016, в независимости от используемой гарнитуры шрифта, напечатанных и отсканированных с разрешением сканера не менее 300 точек на дюйм. Стоит отметить, что в современных сканирующих устройствах показатель разрешения сканера в 300 точек на дюйм соответствует стандартному качеству сканирования.

В ходе проведения третьей группы исследований осуществлена оценка зависимости точности извлечения встроенной информации от величины разрешения изображения, полученного посредством фотографирования, напечатанного на бумаге электронного документа и параметров оформления электронных документов, содержащих текст. Фотографирование напечатанного электронного документа осуществлено с последующей цифровой обработкой сформированного изображения, состоящей из корректировки перспективы изображения и обрезки областей, не относящихся к исходному электронному документу. В качестве используемой метрики оценки точности извлечения информации выступает F-мера.

Результаты точности извлечения встроенной информации из электронных изображений, полученных посредством печати и фотографирования электронных документов, содержащих

текст и оформленных в соответствии с требованиями ГОСТ 7.0.97–2016, представлены на рис. 6.

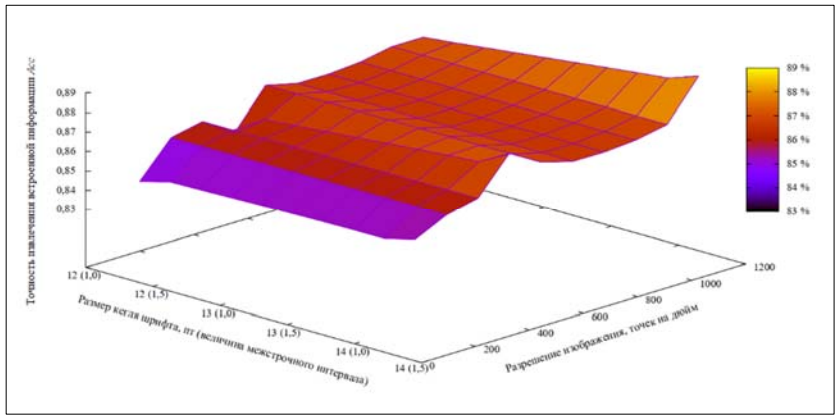


Рис. 6. Зависимость точности извлечения от параметров сфотографированного изображения и оформления текстового документа, содержащего встроенный маркер
Fig.6. Extraction accuracy dependency on the parameters of the photted image and text document design containing an embedded marker

Полученные значения точности извлечения встроенных данных из сфотографированных изображений не позволяет отнести разработанный алгоритм маркирования к алгоритмам с высокой точностью извлечения (близкой или равной 100%) ввиду наличия большого количества ошибок. Указанная особенность не позволяют обеспечить правильное извлечение встроенной информации и точно установить источник утечки информации. Для устранения указанного недостатка может быть применена циклическая схема внедрения маркера, основанная на повторении встраиваемого маркера допустимое число раз совместно с методами помехоустойчивого кодирования.

Точность извлечения разработанного алгоритма маркирования зависит от используемых параметров оформления документов: кегль шрифта и величина межстрочного интервала, а также от разрешения изображения. Точность извлечения разработанного алгоритма маркирования превышает значение в 95 % при следующих параметрах:

- разрешение отсканированного изображения более 300 точек на дюйм, сфотографированного – более 600 точек на дюйм;
- печать и сканирования электронного документа с внедрением закодированного маркера методом помехоустойчивого кодирования;
- печать и фотографирование электронного документа с внедрением закодированного маркера методом помехоустойчивого кодирования по циклической схеме внедрения.

Полученные результаты точности извлечения встроенной информации из изображений, содержащих текстовый документ со встроенным маркером, позволяют перейти к оценке робастности разработанного алгоритма маркирования к осуществлению преобразований и внесению искажений, возникающим в процессе печати, сканирования и фотографирования исходного документа.

2.5 Робастность

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

встроенных данных [24]. Робастность является качественной характеристикой. Проведенная экспериментальная оценка извлекаемости показала, что в разработанном алгоритме маркирования обеспечиваются высокие значения точности извлечения встроенной информации после печати и сканирования исходного документа. Стоит отметить, что процесс сканирования, как и фотографирования, характеризуется наличием искажений и преобразований, которые вносят печатающее, фотосчитывающее устройство или объектив фотоаппаратуры, а также внешние факторы: освещение, вибрация и прочее. Исходя из описанных особенностей, исследование робастности разработанного алгоритма маркирования проводились по двум направлениям. В рамках первого направления исследовалась устойчивость (робастность) встроенной информации к следующим преобразованиям и искажениям, возникающим в процессе печати и сканирования электронного документа:

- поворот изображения;
- изменение соотношения сторон изображения (масштабирование);
- сжатие изображения с потерями (форматы .jpeg, .tif);
- сжатие изображения без потерь (форматы .png, .bmp, .tif, .gif);
- билатеральная фильтрация;
- гауссовская фильтрация (гауссовский фильтр размытия);
- медианная фильтрация;
- внесение фона в изображение.

Особенность разработанного алгоритма заключается в извлечении интервалов между словами, которые расположены на прямой линии относительно центра слов. В случае поворота изображения прямая линия, проходящая через центры слов, преобразуется в ломаную и значения интервалов между словами не могут быть правильно извлечены. Указанная особенность вносит ограничение не только на робастность разработанного алгоритма маркирования к указанному преобразованию, но и на возможность извлечения встроенной информации из анализируемых изображений.

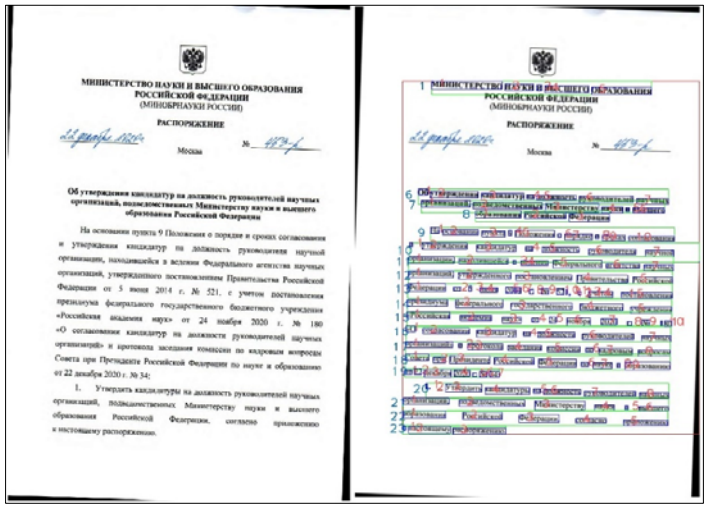


Рис. 7. Извлечение данных из изображения, повернутого на 2 градуса
Fig.7. Extracting data from an image rotated by 2 degrees

В ходе экспериментальной оценки устойчивости разработанного алгоритма к повороту определены предельные значения углов, на которые может быть повернут текст. Разработанный алгоритм позволяет извлекать встроенную информацию из изображения, повернутого не более чем на ±3 градуса. Пример извлечения встроенной информации из изображения, повернутого на 2 градуса, представлен на рис. 7.

В процессе оценки устойчивости разработанного алгоритма к изменению соотношения сторон изображения (масштабирование) проведены эксперименты по растягиванию подписанного изображения как в горизонтальной или вертикальной плоскости, так и в обеих плоскостях одновременно. Диапазон исследуемых значений изменения коэффициента масштабирования составляет 0,5...2,5 (от 1/2 до 2,5 размера исходного документа).

Анализ полученных результатов показал, что разработанный алгоритм маркирования обеспечивает устойчивость встроенной информации к изменению отношения сторон изображения (масштабированию) в пропорциях (коэффициентов масштабирования), находящихся в пределах: 0,5…2,5 относительно исходного размера изображения как в горизонтальной или вертикальной плоскости, так и в обеих плоскостях одновременно.

В хода анализа стойкости разработанного алгоритма маркирования к осуществлению сжатия изображения с потерями, осуществляемое в процессе формирования изображения. В процессе экспериментальной оценки исходное изображение подвергалось сжатию по алгоритму JPEG с показателем качества 10…100 с шагом 10%. В результате анализа полученных результатов установлено, что робастность разработанного алгоритма к сжатию изображения с потерями обеспечивается в случае использования показателя качества сжатия от 30 % и более.

Стойкость разработанного алгоритма к сжатию изображения без потерь заложена в алгоритм маркирования и извлечения встроенного маркера. Так, во время внедрения и извлечения маркера сформированное изображения подвергается растеризации изображения – преобразованию изображения в формат .png, который является одним из представителей форматов изображения, использующим сжатие без потерь или полное отсутствие сжатия. В результате чего можно сделать вывод о том, что разработанный алгоритм обеспечивает устойчивость встроенного маркера к сжатию исходного изображения без потерь.

Экспериментальная оценка стойкости разработанного алгоритма к фильтрации изображения проведена для следующих фильтров: билатеральный, гауссовский (фильтр размытия) и медианный фильтр. Билатеральная фильтрация – нелинейная фильтрация, выполняющая пространственное усреднение в пределах своей маски, применяемая для удаления шума и сглаживания однородных областей.

Исходя из особенностей применения фильтра, направленных на сглаживание областей изображения, результат извлечения встроенной информации обладает устойчивостью к применению данного фильтра. Предельные значения примененного фильтра составляют: маска фильтра (диаметр соседних пикселей) – не более 10, сигма-фильтра в цветной и пространственной областях – не более 150.

Гауссовский (гауссов) фильтр – фильтр линейного сглаживания, предназначенный для удаления шума, описываемого гауссовским (нормальным) законом распределения. Как и билатеральный гауссовский фильтр применяется для уменьшения резких изменений в градациях серого изображения.

Полученные значения позволяют сделать вывод об устойчивости разработанного алгоритма маркирования к осуществлению гауссовской фильтрации изображения с предельным значением ядра свертки в (30, 30) по высоте и ширине или радиуса размытия в 8 пикселей.

Медианный фильтр представляет собой эвристический метод обработки изображения, который удаляет из сигнала (изображения) фрагменты с размерами, меньшими, чем половина размера окна фильтра, и при этом мало искажает или почти совсем не искажает остальные

участки сигнала (изображения). Оценка устойчивости к медианной фильтрации изображения, содержащего текстовые данные, производилась за счет применения медианного фильтра с шагом изменения радиуса размытия от 1 до 9 пикселей. Предельное значение радиуса размытия 9 характеризует пороговое значение робастности к медианной фильтрации разработанного алгоритма маркирования.

Помимо фильтрации формируемое изображение может подвергаться внесению дополнительных элементов, в частности изменение цвета фона изображения, вызванное износом сканирующего устройства или подкладыванием дополнительных листов бумаги, обладающих оттенками цвета отличным от белого.

Полученные результаты извлечения встроенной информации после внесения дополнительно фона изображения позволяют сделать вывод о наличии стойкости разработанного алгоритма маркирования к указанному типу искажения. Граничные значения стойкости определяются прозрачностью фона. В случае использования прозрачности фона меньше 10% исходный текст и графические изображения в документе будут нечитаемыми. Указанный факт ограничивает стойкость разработанного алгоритма маркирования предельным значением прозрачности фона изображения не менее чем 10%.

В результате проведенной экспериментальной оценки робастности разработанного алгоритма маркирования к осуществлению преобразования формата, обусловленного печатью и сканированием, получены результаты, представленные в табл. 6.

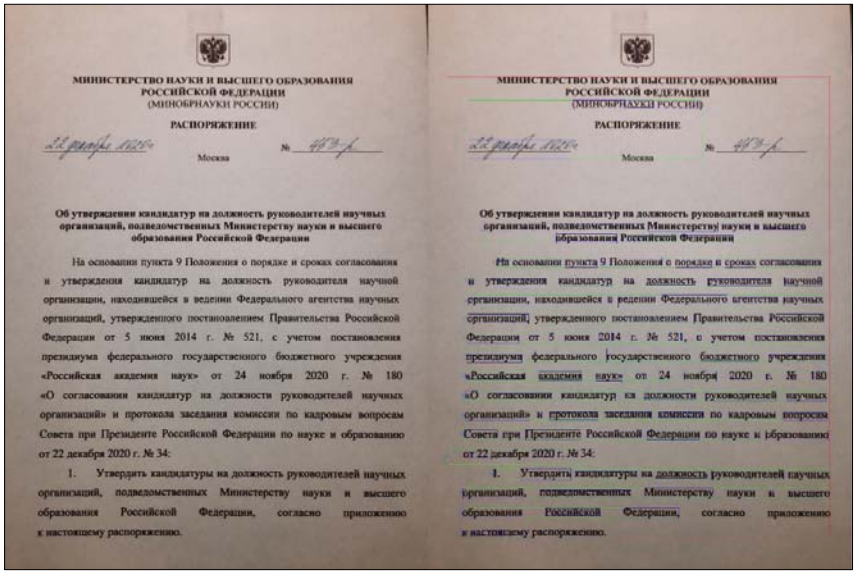
Табл. 6. Робастность разработанного алгоритма к преобразованиям и искажениям, возникающим в процессе печати и сканирования
Table 6. Robustness of the developed algorithm to transformations and distortions that occur during printing and scanning

Тип преобразования	**Параметры робастности**
Поворот изображения	±3∘ относительно горизонтально выравненного текста
Изменение соотношения сторон	Коэффициент масштабирования 0,5…2,5 относительно исходного размера
Сжатие изображения с потерями	По алгоритму JPEG с показателем качества не менее 30 %
Сжатие изображения без потерь	Алгоритмы: RLE, LZW и Deflate
Балатеральная фильтрация	Диаметр соседних пикселей не более 10, сигма-фильтра в цветной и пространственной областях не более 150
Гауссовская фильтрация	Размер ядра по высоте и ширине не более (30, 30)
Медианная фильтрация	Предел ядра свертки 9 пикселей

Результаты робастности, представленные в табл. 6, а также значения точности извлечения встроенной информации из изображений, содержащих встроенный маркер, позволяют сделать вывод о возможности гарантированного извлечения встроенного маркера (с точностью извлечения более 95%) из изображений, полученных посредством печати и сканирования электронных документов.

В рамках второго направления исследований осуществлена оценка устойчивости (робастности) встроенной информации к преобразованию электронного документа в изображения посредством печати и фотографирования, а также осуществления сопутствующих искажений. Отличительной особенностью процесса фотографирования от сканирования является наличие геометрических искажений в трех плоскостях: горизонтальный наклон, вертикальное отклонение и вращение относительно центра. Кроме того, процесс фотографирования характеризуется наличием искажений, вносимых объективом фотоаппаратуры и условиями съемки: неравномерно освещенные области, муар, виньетирование, искажение перспективы и т.д.

Указанные особенности требуют проведения дополнительного этапа предварительной обработки изображения, направленного на коррекцию перспективы и обрезку областей изображения, изначально не относившихся к исходному текстовому документу. Без проведения этапа предварительной обработки извлечение встроенной информации не может быть реализовано с требуемым значением точности извлечения. Пример извлечения информации из изображения, содержащего встроенные данные, сформированного посредством фотографирования, напечатанного на бумаге электронного документа, содержащего текст, представлен на рис. 8.


МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
(МИНОБРНАУКИ РОССИИ)

РАСПОРЯЖЕНИЕ

Москва

Об утверждении кандидатур на должность руководителей научных организаций, подведомственных Министерству науки и высшего образования Российской Федерации

На основании пункта 9 Положения о порядке и сроках согласования и утверждения кандидатур на должность руководителя научной организации, находившейся в ведении Федерального агентства научных организаций, утвержденного постановлением Правительства Российской Федерации от 5 июня 2014 г. № 521, с учетом постановления президиума федерального государственного бюджетного учреждения «Российская академия наук» от 24 ноября 2020 г. № 180 «О согласовании кандидатур на должности руководителей научных организаций» и протокола заседания комиссии по кадровым вопросам Совета при Президенте Российской Федерации по науке и образованию от 22 декабря 2020 г. № 34:

1. Утвердить кандидатуры на должность руководителей научных организаций, подведомственных Министерству науки и высшего образования Российской Федерации, согласно приложению к настоящему распоряжению.

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
(МИНОБРНАУКИ РОССИИ)

РАСПОРЯЖЕНИЕ

Москва

Об утверждении кандидатур на должность руководителей научных организаций, подведомственных Министерству науки и высшего образования Российской Федерации

На основании пункта 9 Положения о порядке и сроках согласования и утверждения кандидатур на должность руководителя научной организации, находившейся в ведении Федерального агентства научных организаций, утвержденного постановлением Правительства Российской Федерации от 5 июня 2014 г. № 521, с учетом постановления президиума федерального государственного бюджетного учреждения «Российская академия наук» от 24 ноября 2020 г. № 180 «О согласовании кандидатур на должности руководителей научных организаций» и протокола заседания комиссии по кадровым вопросам Совета при Президенте Российской Федерации по науке и образованию от 22 декабря 2020 г. № 34:

1. Утвердить кандидатуры на должность руководителей научных организаций, подведомственных Министерству науки и высшего образования Российской Федерации, согласно приложению к настоящему распоряжению.

Рис. 8. Извлечение данных из изображения, сформированного посредством фотографирования
Fig.8. Extracting data from an image formed by photographing

Оценка устойчивости встроенной информации к искажениям, возникающим в процессе печати и фотографирования электронного документа, проводилась по тем же критериям, что и оценка устойчивости к искажениям, возникающим при печати и сканировании. Разработанный алгоритм показал те же значения робастности применении операций печать и фотографирование, что и значения, представленные в табл. 6. При этом точность извлечения встроенной информации не превышает значение в 89 %. Полученные значения робастности к сканированию и фотографированию позволяют описать робастность разработанного алгоритма маркирования следующим образом:

- печать и сканирование электронного документа: робастность к искажениям, представленным в табл. 6, при использовании методов помехоустойчивого кодирования в процессе маркирования;
- печать и фотографирование электронного документа: наличие дополнительного этапа обработки изображения, направленного на корректировку перспективы изображения и удаления областей, не присутствовавших в исходном текстовом документе; использование фотоаппаратуры, формирующей изображение не менее чем 4000 × 3000 (12 мегапикселей); использование циклической схемы встраивания и методов помехоустойчивого (при наличии достаточной емкости встраивания).

4. Заключение

Проведенная экспериментальная оценка основных параметров разработанного алгоритма маркирования, основанного на изменении интервалов между словами, позволяет сделать вывод о повышении защищенности электронных документов, содержащих текстовую информацию, от утечки посредством печати, сканирования или фотографирования с последующей отправкой изображения в случае внедрения предложенного подхода в компоненты DLP- или SIEM-системы. При этом наличие одиночных ошибок в процессе извлечения информации из отсканированных изображений и низкие значения точности извлечения встроенных данных из сфотографированных изображений требуют дальнейшего совершенствования предложенного подхода к маркированию электронных документов, а также проведения сравнительного анализа полученных результатов с существующими аналогами. Решение поставленных задач является направлением дальнейших исследований.

Список литературы / References

[1] Cybersecurity threatscape: Year 2021 in review. Positive Technologies, 2022, 23 p. Available at: https://www.ptsecurity.com/upload/corporate/ww-en/analytics/Cybersecurity_threatscape_2021_ENG.pdf, accessed 10.08.2022.

[2] Отчёт об исследовании утечек информации ограниченного доступа в 2021 году. InfoWatch, 2022 г., 32 стр. / Restricted Information Leakage Study Report in 2021. InfoWatch. 2021, 32 p. Available at: https://www.infowatch.ru/analytics/analitika/v-2021-stalo-bolshe-umyshlennykh-utechek, accessed 10.08.2022 (in Russian).

[3] Jain M., Lenka S.K. A Review on Data Leakage Prevention using Image Steganography. International Journal of Computer Science Engineering, vol. 5, no 2, 2016, pp 56-59.

[4] Lopez G., Richardson N., Carvajal J. Methodology for Data Loss Prevention Technology Evaluation for Protecting Sensitive Information. Revista Politecnican, vol. 36, no 3, 2015, pp. 60-69.

[5] Alneyadi S., Sithirasenan E., Muthukkumarasamy V. A survey on data leakage prevention systems. Journal of Network and Computer Applications, vol. 62, 2016, pp. 137-152.

[6] Jadhav P., Chawan P.M. Data Leak Prevention system: A Survey. International Research Journal of Engineering and Technology, vol. 6, no. 10, 2019, pp. 197-199.

[7] Козачок А.В., Копылов С.А. и др. Алгоритм маркирования текстовых документов на основе изменения интервалов между словами, обеспечивающий устойчивость к преобразованию формата. Труды ИСП РАН, том 5, вып. 5, 2021 г., стр. 131-146. DOI: 10.15514/ISPRAS-2021-33(4)-10. / Kozachok A.V., Kopylov S.A. et al. Text documents marking algorithm based on interword distances shifting invariant to format conversion. Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 4, 2021, pp. 131-146 (in Russian).

[8] Kozachok A.V., Kopylov S.A. et al. Text marking approach for data leakage prevention. Journal of Computer Virology and Hacking Techniques, vol. 15, no. 3, 2019, pp. 219-232.

[9] Salomon D. Data privacy and security: encryption and information hiding. Springer Science & Business Media, New York, 2003. 469 p.

[10] Kapila B., Thind T. Review and analysis of data security using image steganography. In Proc. of the 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), 2021, pp.227-231.

[11] Woo C.-S. Digital image watermarking methods for copyright protection and authentication. PhD Thesis. Information Security Institute, Faculty of Information Technology, Queensland University of Technology, 2007, 197 p.

[12] Mohanarathinam A., Kamalraj S. et al. Digital watermarking techniques for image security: a review. Journal of Ambient Intelligence and Humanized Computing, vol. 11, 2020, pp. 3221-3229.

[13] Khadam U., Iqbal M.M. et al. Digital Watermarking Technique for Text Document Protection Using Data Mining Analysis. IEEE Access, vol. 7, 2019, pp. 64955-64965.

[14] Kozachok A.V., Kopylov S.A. Estimation of Watermark Embedding Capacity with Line Space Shifting. In Proc. of the Ivannikov Memorial Workshop (IVMEM), 2020, pp. 29-34.

[15] Национальный стандарт Российской Федерации. Система стандартов по информации, библиотечному и издательскому делу. Организационно-распорядительная документация. Требования к оформлению документов. ГОСТ Р 7.0.97–2016, Стандартинформ, 2019 г., 32 стр. /

National standard of the Russian Federation, System of standards on information, librarianship and publishing. Organizational and administrative documentation. Requirements for presentation of recordsю GOST R 7.0.97–2016, Standartinform, 2019, 32 p. (in Russian).
[16] Zhou N.R., Hou W.M.X., Wen R.H. Imperceptible digital watermarking scheme in multiple transform domains. Multimedia Tools and Applications, vol. 77, 2018, pp. 30251–30267.
[17] Wu J.Y., Huang W.L., Xia-Hou W.M. Imperceptible digital watermarking scheme combining 4-level discrete wavelet transform with singular value decomposition. Multimedia Tools and Applications, vol. 79, 2020, pp. 22727–22747.
[18] Грибунин В.Г., Оков И.Н., Туринцев И.В. Цифровая стеганография. Москва, СОЛОН-Пресс, 2017 г., 262 стр. / Gribunin V.G., Okov I.N., Turincev I.V. Digital steganography. Moscow, SOLON-Press, 2017, 262 p. (in Russian).
[19] Коржик В.И. Цифровая стеганография и цифровые водяные знаки. Санкт-Петербург, СПбГУТ, 2017 г., 424 стр. / Korzhik V.I. Digital Steganography and Digital Watermarking. Saint-Petersburg, SPbSUT, 2016, 226 p. (in Russian)
[20] Козачок А.В., Копылов С.А., Бочков М.В. Оценка параметров необнаруживаемости разработанного подхода к маркированию текстовых электронных документов. Вопросы кибербезопасности, no. 1(35), 2020, стр. 62-73 / Kozachok A.V., Kopylov S.A., Bochkov M.V. Undetectability Parameters Estimation of the Developed Approach to Text Electron Documents Marking. no 1(35), 2020, pp. 62-73 (in Russian).
[21] Karampidis K., Kavallieratou E., Papadourakis G. A review of image steganalysis techniques for digital forensics. Journal of Information Security and Applications, vol. 40, 2018, pp. 217-235.
[22] Yang Z., Huang Y., Zhang Y.-J. A fast and efficient text steganalysis method. IEEE Signal Processing Letters, vol. 26, no. 4, 2019, pp. 627-631.
[23] Kadian P., Arora S.M., Arora N. Robust Digital Watermarking Techniques for Copyright Protection of Digital Data: A Survey. Wireless Personal Communications, vol. 118, 2021, pp. 3225-3249.
[24] Menendez-Ortiz A., Feregrino-Uribe C. et al. A Survey on Reversible Watermarking for Multimedia Content: A Robustness Overview, IEEE Access, vol. 7, 2019, pp. 132662-132681.

Информация об авторах / Information about authors

Александр Васильевич КОЗАЧОК – доктор технических наук, доцент, заведующий лабораторией безопасного программного обеспечения и анализа данных. Сфера научных интересов: методы и системы защиты информации, кибербезопасность, машинное обучение, анализ данных.

Alexander Vasilievich KOZACHOK – Doctor of Technical Sciences, associate professor, Head of the Laboratory of Secure Software and Data Analysis. Research interests: algebraic structures in the information security methods and systems, cybersecurity, machine learning, data analysis.

Василий Иванович КОЗАЧОК – доктор социологических наук, профессор, сотрудник Академии Федеральной службы охраны Российской Федерации. Его научные интересы включают: безопасность информации, защита информации от несанкционированного доступа, построение информационных систем в защищённом исполнении.

Vasilii Ivanovich KOZACHOK – Doctor of Sociological Sciences, Professor. Employer of the Academy of Federal Guard Service. His research interests include: information security, information unauthorized access protection, information systems construction in a secure design.

Сергей Александрович КОПЫЛОВ – кандидат технических наук, сотрудник Академии Федеральной службы охраны Российской Федерации. Его научные интересы включают: методы машинного обучения, обработка цифровых изображений, текстовая стеганография.

Sergey Alexandrovich KOPYLOV – PhD in Technical Sciences. Employer of the Academy of Federal Guard Service. His research interests include machine learning methods, digital image processing, text steganography.

Павел Николаевич ГОРБАЧЕВ – сотрудник Академии Федеральной службы охраны Российской Федерации. Его научные интересы включают: информационная безопасность,

методы машинного обучения, распознавание образов, текстовая стеганография, обработка изображений.

Pavel Nikolaevich GORBACHEV is employer of the Academy of Federal Guard Service. His research interests include: information security, machine learning methods, pattern recognition text steganography and image processing.

Юрий Витальевич МАРКИН – кандидат технических наук, научный сотрудник. Область научных интересов: информационная безопасность, анализ сетевого трафика, обработка изображений, алгоритмы машинного обучения.

Yury Vital'evich MARKIN – PhD in Technical Sciences. Researcher. Scientific interests: information security, network traffic analysis, image processing, machine learning algorithms.

Дмитрий Олегович ОБЫДЕНКОВ – аспирант. Его научные интересы включают методы сокрытия и защищённой передачи информации, компьютерные сети, технологии анализа сетевого трафика.

Dmitry Olegovich OBYDENKOV is a graduate student. His scientific interests include methods for information hiding and secure transmission, computer networks, technologies of network traffic analysis.