УДК 004.912+004.932.2

А.В. Козачок, С.А. Копылов

Подход к маркированию текстовых электронных документов и его сравнение с существующими аналогами

Обеспечение безопасности текстовых данных, представленных как в электронном, так и в напечатанном виде, является одним из актуальных направлений исследований. В данной работе представлены результаты экспериментальной оценки разработанного подхода к маркированию текстовых документов, рассмотрены особенности формируемого водяного. Проведен обзор существующих исследований в области маркирования текстовых данных, определены их основные достоинства и недостатки. Представлены результаты сравнительного анализа параметров встраивания и извлечения разработанного подхода с рассмотренными аналогами. Определены направления дальнейших исследований.

Ключевые слова: защита информации, маркирование текстовых данных, текстовая стеганография.
doi: 10.21293/1818-0442-2019-22-2-52-60

В последние десятилетия резко возросло количество инцидентов нарушения информационной безопасности в области защиты интеллектуальной собственности и защиты авторских прав. Основным объектом защиты современных средств обеспечения безопасности данных являются изображения и данные мультимедиа. В то же время проблеме защиты авторских прав владельцев текстовой информации и защиты текстовой информации от утечки не уделяется должного внимания [1].

Современные средства защиты текстовых данных от утечки и защиты авторских прав в должной мере не позволяют обеспечить надежную защиту [2–6]. Указанный факт обусловлен возможность преобразования формата исходного текстового документа в изображение, содержащее текст. При этом обнаружение исходных данных в существующих системах защиты основано на средствах оптического распознавания символов, характеризующихся наличием ошибок в процессе распознавания. Для устранения указанного недостатка необходимо разрабатывать новые методы защиты текстовой информации от утечки. В качестве такого метода может выступать подход к маркированию текстовых данных, основанный на стеганографическом внедрении робастного водяного знака.

Подход к маркированию электронных текстовых документов

В качестве подхода к маркированию текстовых данных выступает подход, разработанный авторами, описанный в [7–10]. Маркирование осуществляется посредством изменения величины межстрочного интервала на величину перцептивно невидимую для человеческого глаза (рис. 1). В качестве маркера может выступать идентификационная информация, характеризующая владельца данных, сами данные, либо другая метаинформация. Встраивание маркера, представленного робастным водяным знаком (РВЗ), реализовано следующим образом:

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

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

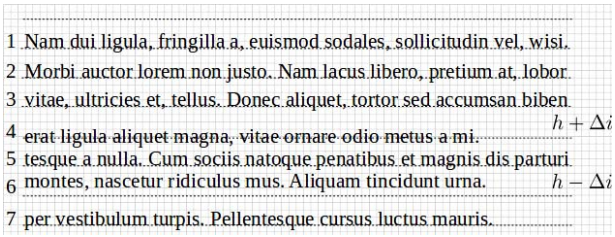

Рис. 1. Вариант встраивания информации на основе изменения межстрочного интервала

В ходе экспериментальной оценки предложенного подхода установлены следующие параметры:

1. Емкость встраивания зависит от кегля шрифта и величины межстрочного интервала исходного текста и не зависит от гарнитуры и параметров используемого шрифта.

Значения предельно достижимой емкости встраивания представлены в табл. 1.

Таблица 1

Предельно достижимая емкость встраивания

Кегль шрифта (пт)	Межстрочный интервал (множитель)	Предельно достижимая емкость встраивания (бит)
10	1	60
10	1,25	48
10	1,5	40
12	1	49
12	1,25	39
12	1,5	33
14	1	42
14	1,25	33
14	1,5	28

2. Граница перцептивной невидимости встроенных данных находится в пределах изменения величины межстрочного интервала на $\pm 0{,}15$ от исходного.

3. Точность извлечения встроенных данных из изображений, содержащих текст, составляет не ме-

нее 95% при использовании параметра встраивания, характеризующимся увеличением величины множителя межстрочного интервала на 0,04 и более с шагом 0,01. Указанные значения точности извлечения соответствуют изображениям с показателем разрешения не менее 150 точек на дюйм. Результат извлечения встроенной информации из изображения, содержащего текст с размером шрифта 14 пт, величиной межстрочного интервала 1 и величиной изменения межстрочного интервала 0,10 интервала (0,49 мм), представлен в табл. 2.

В процессе экспериментальной оценки точности извлечения данных было извлечено более 10 000 бит (более 250 страниц текстовой информации), что позволяет утверждать о том, что доверительный интервал равен 0,95 при точности 0,01. Результаты оценки зависимости точности извлечения данных, ошибок первого и второго рода от величины изменения межстрочного интервала из изображений, содержащих РВЗ, представлены на рис. 2.

Разработанный подход обеспечивает инвариантность встроенных данных к следующим преобразованиям:

– преобразование формата электронного текстового документа в текстовое изображение, в том числе посредством операции «печать–сканирование»;

– поворот текстового изображения на любой угол;

– масштабирование текстового изображения со значением множителя масштабирования, не превышающим ±1,5;

– фильтрация текстового изображения (медианная, гауссовская, модовая);

– сжатие с потерями при использовании показателя качества, не превышающего 50%.

Таблица 2

Результат извлечения информации из изображения, содержащего текст

Разрешение изображения (DPI)	Время обработки, с	Число строк в исходном документе	Извлеченное число строк	Точность	Вероятность ложных срабатываний	Вероятность пропуска цели
25	0,9	41	40	0,667	0,033	0,30
50	4	41	41	0,886	0,033	0,081
100	15	41	41	0,984	0,016	0
150	34	41	41	0,992	0,008	0
200	62	41	41	0,992	0,008	0
250	105	41	41	0,992	0,008	0
300	166	41	41	0,992	0,008	0

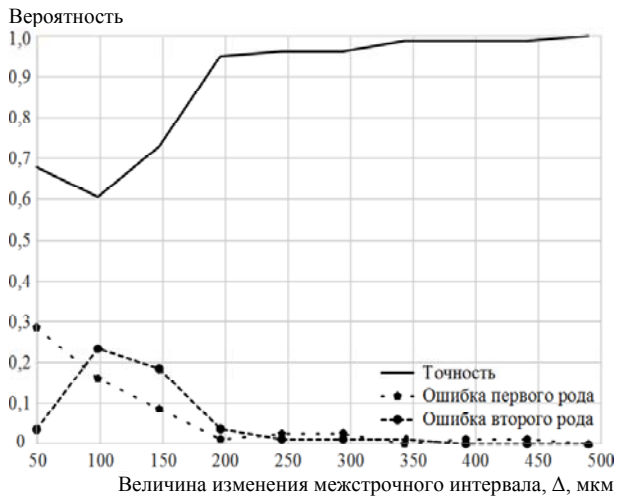

Рис. 2. Оценка зависимости точности извлечения информации, ошибок первого и второго рода от величины изменения межстрочного интервала

Полученные результаты экспериментальной оценки разработанного подхода к маркированию текстовых данных позволяют сделать вывод о возможности применения указанного подхода для идентификации текстовых данных, защиты от утечки и защиты авторского права. В то же время необходимо провести сравнительную оценку разработанного подхода с существующими аналогами в области маркирования текстовых данных.

Обзор текущих исследований в области маркирования текстовых данных

В ходе проведения сравнительного анализа исследований в области маркирования текстовых данных рассмотрены следующие подходы предложенные: Брассилом [11–13], Коксом [14], Лиу [15], Хуангом [16] и Алаттаром [17]. В качестве базового аналога выступает подход, описанный Брассилом.

В работах Брассила [11–13] представлен подход к маркированию текстовых документов, основанный на внедрении кодовой последовательности посредством смещения линии текста. Кодирование информации осуществляется следующим образом:

– линия текста нечетной строки абзаца не изменяется, что соответствует встраиваемому «0»;

– линия текста четной строки абзаца может быть смещена вверх, что соответствует встраиваемой «–1» и вниз, что соответствует встраиваемой «+1»;

– величина абзацного отступа не изменяется.

Емкость встраивания рассматриваемого подхода зависит от размера кегля гарнитуры шрифта и ограничена величиной 19 бит (при использовании кегля размером 12 пт). Указанные значения емкости встраивания обусловлены требованиями, предъявляемыми к решающему правилу детектора в процессе извлечения встроенной информации.

Извлечение встроенной информации из изображений, содержащих исходный текст, состоит из следующих этапов:

– предварительная обработка изображения;

– выравнивание изображения по направлению текста;

– построение горизонтального профиля проекции изображения;

– определение направления смещения строки:

1) измерение расстояния между базовыми линиями профиля проекции;

2) измерение расстояния между центроидами профиля проекции.

В ходе предварительной обработки из изображения удаляется шум, вносимый принтером в процессе печати документа. Полученное изображение выравнивается по направлению текста, после чего осуществляется построение горизонтального профиля проекции обработанного изображения. Профиль проекции представляет собой график суммирования интенсивностей пикселей каждой строки изображения. Для определения положения смещения линии текста используются два подхода. Первый подход основан на измерении расстояния между базовыми линиями профиля проекции. В качестве базовой линии выступает пик профиля проекции линии текста, наиболее близко расположенный относительно дна. В отличие от базовой линии центроид представляет собой центр масс пика профиля проекции линии текста. В случае использования подхода к извлечению линии текста, основанного на измерении расстояния между базовыми линиями, решающее правило детектора имеет следующий вид:

если $s_{i-1} > s_i$: базовая линия смещена вниз,
если $s_{i-1} < s_i$: базовая линия смещена вверх,
иначе: смещение не определено,

где s_{i-1} – расстояние между линиями $i-1$ и i; s_i – расстояние между линиями i и $i+1$ (при этом $i-1$ и $i+1$ – линии, положение которых не изменялось в процессе встраивания).

Решающее правило детектора, используемого в подходе к извлечению линии текста, основанного на расстоянии между центроидами, описывается следующим образом:

если $s_{i-1} - t_{i-1} > s_i - t_i$: базовая линия смещена вниз,
иначе: базовая линия смещена вверх,

где s_{i-1} и s_i – расстояние между центроидами линий $i-1$ и i и линий i и $i+1$, соответственно, текста, содержащего встроенные данные; t_{i-1} и t_i – расстояние между центроидами линий $i-1$ и i и линий i и $i+1$ соответственно исходного текста.

Предложенные подходы к определению смещения строки позволяют повысить точность извлечения встроенной информации. Однако, использование в процессе анализа двух строк для извлечения одного бита встроенной информации снижает предельно достижимую емкость встраивания в два раза. Кроме того, для правильного детектирования и извлечения встроенной информации необходимо наличие как исходного изображения, так и опорного базиса (эталонного образца), позволяющих определить направление смещения линии текста.

В ходе экспериментальной оценки установлено, что минимальная величина сдвига линии текста ограничена величиной в 1 пиксель текста, выводимого на печать с разрешением 300 точек на дюйм (примерно 0,085 мм). Данное ограничение обусловлено разрешающей способностью принтера, на котором осуществляется печать документа. При этом данный подход характеризуется следующими значениями предельно достижимой емкости встраивания: кегль размером 8 пт – 23 бита; 10 пт – 21 бит и 12 пт – 19 бит.

Внедрение кодовой последовательности осуществлялось в страницу текста, набранного гарнитурой Times New Roman с размером кегля 8, 10 и 12 пт посредством смещения линии текста вверх и вниз на величину 1, 2 и 3 пикселя.

В ходе извлечения встроенной информации подход, основанный на измерении расстояния между центроидами, позволил безошибочно установить смещение линии текста для каждой комбинации встраивания. При этом подход, основанный на измерении расстояния между базовыми линиями профиля проекции, характеризуется наличием ошибок. Так, точность извлечения данных для кегля размером 8 пт и величиной встраивания 1 пиксель составляет 78% (5 строк не извлечены), а для 12 пт и 1 пиксель – 94%.

Подход, основанный на измерении расстояния между центроидами, позволил безошибочно определить смещение линии текста кеглем 10 пт. В то же время точность извлечения подхода, основанного на измерении расстояния между базовыми линиями, составила не более 76%.

Представленный подход к маркированию текстовых документов характеризуется следующими параметрами робастности встроенных данных:

– преобразование формата электронного текстового документа в изображение, содержащее текст, посредством применения операции «печать–сканирование»;

– перекопирование (до 10 раз) напечатанного текстового документа;

– поворот изображения на угол $\pm 3°$;

– размытие изображения (гауссовская фильтрация);

– масштабирование изображения, содержащего встроенные данные с коэффициентом масштабирования $\pm 4\%$.

Маркирование текстовых документов, предложенное в работе [14], основано на технологии преобразования доменов. Цифровой водяной знак (ЦВЗ) представляет собой информацию в виде последовательности действительных чисел $X = x_1, x_2, \ldots, x_N$ длиной N. ЦВЗ встраивается в частотные компоненты изображения, содержащего текстовые данные, следующим образом:

– вычисление частотных коэффициентов АС изображения $D(x,y)$, содержащего текст, $C_D(k_x, k_y)$, посредством двумерного дискретного косинусного преобразования;

– извлечение последовательности $V = v_1, v_2, \ldots, v_N$, состоящей из полученных коэффициентов $C_D(k_x, k_y)$, значение которых превышает N;

– сложение элементов последовательности V с ЦВЗ X по следующему правилу: $v_i' = v(i + \alpha \cdot x_i)$, где α – параметр масштабирования, определяющий степень искажения изображения;

– встраивание последовательности V' в $C_D(k_x, k_y)$ вместо V посредством обратного дискретного косинусного преобразования.

В результате встраивания формируется подписанное текстовое изображение D'. К сформированному изображению D' могут быть применены различные операции обработки, в том числе передача по каналу связи, в ходе которых в текстовое изображение могут быть внесены искажения и шум. В результате таких преобразований на анализатор ЦВЗ поступает преобразованное изображение D^*, содержащее встроенные данные. Для правильного извлечения встроенных данных на стороне приема необходимо наличие исходного изображения и соответствующего ему ЦВЗ. Процесс извлечения встроенных данных состоит из следующих этапов:

– извлечение последовательностей $V^* = v_1^*, v_2^*, \ldots, v_N^*$ и $V = v_1, v_2, \ldots, v_N$ из частотных коэффициентов изображений D^* и D соответственно посредством дискретного косинусного преобразования;

– вычисление искаженного ЦВЗ $X^* = V^* - V$;

– сравнение исходного X и искаженного X^* ЦВЗ согласно указанному правилу:

$$\mathrm{sim}(X, X^*) = \frac{X^* \cdot X}{\sqrt{X^* \cdot X^*}};$$

– вывод об аутентичности встроенных данных.

В ходе экспериментальной оценки емкости встраивания разработанного подхода было установлено, что длина водяного знака может быть произвольной и зависит от количества используемых частотных компонент, в которые осуществляется встраивание. Так, в случае четырехкратного увеличения числа используемых частотных компонент вдвое увеличивается предельно достижимая емкость встраивания. Однако такое увеличение в значительной степени искажает качество исходного изображения, что приводит к отсутствию перцептивной невидимости встроенных данных. Кроме того, в процессе встраивания информации в частотные компоненты изображения осуществляется изменение фона изображения. Указанные особенности не позволяют отнести рассмотренный подход к перцептивно невидимым подходам встраивания при использовании ЦВЗ, характеризующихся большой длиной.

В рассматриваемом подходе предложено использовать ЦВЗ длиной, не превышающей 1000 символов, что соответствует наличию незначительных изменений в фоне подписанного изображения, не влияющих на невидимость встроенных данных к визуальному анализу.

Оценка извлекаемости и робастности разработанного подхода производилась посредством извлечения встроенной информации как после применения различных преобразований и внесения искажений, так и без таковых. Извлекаемость ЦВЗ основана на способности детектора обнаружить встроенный ЦВЗ среди случайно сгенерированных. Практические результаты показали, что среди 1000 случайно сгенерированных ЦВЗ только в 1 случае детектор обнаружил встроенные данные. При этом количественная оценка точности извлечения встроенных данных из изображений не представлена.

В ходе оценки робастности подхода к маркированию текстовых данных [13] учтены результаты экспериментов, проведенных в [14]. Разработанный ЦВЗ устойчив к следующим преобразованиям:

– внесение в изображение гауссовского шума;

– масштабирование изображения с множителем 0,5;

– обрезка не более 75% изображения;

– сжатие изображения по алгоритму JPEG с показателем качества не более 10%.

Представленный подход характеризуется следующими ограничениями:

– низкая перцептивная невидимость при большой длине ЦВЗ;

– низкая робастность к преобразованию формата документа, обусловленная операцией «печать–сканирование» и фотокопированием;

– отсутствие устойчивости к вращению, сдвигу и повороту изображения;

– отсутствие устойчивости к фильтрации (медианной, усредненной);

– вычислительная сложность извлечения встроенных данных.

В работе [15] предложена гибридная схема маркирования текстовых документов. В качестве алгоритма встраивания ЦВЗ выступает один из подходов, описанных в [11–13]. Помимо подхода к встраиванию, основанного на смещении положения линии текста, авторами [11–13] предложено использовать смещение слова внутри строки текста, а также кодирование символов.

Для извлечения встроенных данных использован подход, предложенный в [14]. Отличительной особенностью разработанного подхода от [14] является процедура очистки фона подписанного изображения. Применение данной технологии позволяет увеличить емкость встраивания в 10 раз по сравнению с [14], при этом встроенные данные находятся в границах перцептивной невидимости.

В то же время использование алгоритма встраивания, основанного на сдвиге линии или на смещении слова, характеризуется наличием ошибок в процессе извлечения данных. Кроме того, для корректного извлечения данных детектору необходимо наличие исходного изображения.

В качестве подхода к маркированию текстовых данных может быть рассмотрен подход к внедрению ЦВЗ в текстовые изображения, описанный в [16]. Для маркирования текстового документа предложе-

но использовать свободное пространство (пространство пробелов) между словами. Кодирование информации реализуется посредством преобразования пустого текстового пространства, характеризующегося средним расстоянием между словами в строке, в вид синусоидальной формы. Встраивание информации в изображение, содержащее текст, реализовано следующим образом:

– определение ключевой строки текста (строка текста с заданным количеством слов);

– формирование массива строк S_w посредством выбора равных или превышающих ключевую строку по количеству слов;

– вычисление среднего значения пустого пространства строки S_a и текста a для сформированного массива строк:

$$a = \frac{\sum_{m=u}^{v} S_{am}}{v-u+1},\ 0 \le u < v < N, \qquad (1)$$

где u,v – первая и последняя линия (строка) сформированного массива; m – индекс строк текста сформированного массива; S_{am} – среднее значение пустого пространства строки;

– формирование ЦВЗ для каждой строки массива согласно выражению

$$W_m = C \cdot a \cdot \sin(\omega(m-u)+\phi),$$

где ω – частота; ϕ – начальная фаза; C – константа, определяющая амплитуду синусоидальной волны (диапазон значений 0,2–0,3);

– в каждой линии массива строк S_a заменяется на S_{am} следующим образом:

$$S'_{am} = a + W_m, \text{ если } m \in S_w;$$

– корректировка величины интервала между словами посредством расширения (сжатия) пустого пространства после каждого слова из массива строк S_w в синусоидальный вид.

Извлечение встроенной информации реализуется следующим образом:

– восстановление массива строк S_w посредством ключевой строки текста;

– вычисление среднего значения пустого пространства текста посредством (1);

– извлечение ЦВЗ W_m из подписанного изображения посредством вычитания a;

– определение начальной фазы за счет вычисления взаимной корреляции полученного сигнала синусоидальной формы.

В ходе экспериментальной оценки разработанного алгоритма были протестированы 6 изображений, содержащих текст, со следующими параметрами:

– разрешение изображения 300 точек на дюйм;

– кегль шрифта 11 пт;

– выравнивание по ширине текста.

В ходе оценки извлекаемости разработанного подхода к маркированию текстовых данных произведена оценка емкости встраивания и точности извлечения встроенных данных из изображений, содержащих текст. Емкость встраивания характеризуется произвольной длиной ЦВЗ и зависит от формы синусоидальной волны, параметров, которыми она описывается, а также от параметров ключевой строки. Так, на примере тестового изображения, содержащего 52 строки, емкость встраивания варьируется в пределах от 15 до 23 бит. При этом предельно достижимая емкость встраивания для текста в 52 строки, набранного кеглем размером 11 пт, составляет 42 бита.

Проведение 4 экспериментов по извлечению встроенных данных не позволяет произвести количественную оценку в пределах доверительного интервала 0,01. Однако результат извлечения характеризуется одиночными ошибками извлечения. Так, из встроенных в текстовые данные 68 бит информации правильно извлечены 64 бита. Указанная особенность обусловлена алгоритмом встраивания, а именно процессом расширения (сжатия) строк. Расширение (сжатие) строк приводит к уменьшению расстояния между словами, вплоть до наложения слов друг на друга, что в процессе извлечения не позволяет правильно обнаружить и извлечь встроенную информацию.

Перцептивная невидимость встроенных данных достигается за счет выбора ключевой строки с количеством строк, близким к максимальному значению, а также небольших значений константы C, определяющей амплитуду синусоидальной волны.

Разработанный подход обеспечивает робастность к следующим преобразованиям:

– наклон (поворот) текстового изображения на угол не более 5°;

– передискретизация текстового изображения с параметрами качества не ниже 50% от исходного изображения при применении фильтра низких частот;

– преобразование формата текстового изображения посредством применения операции «печать–сканирование»;

– перекопирование напечатанного текстового изображения (до 10 раз).

Для правильного извлечения встроенной информации из изображений, содержащих РВЗ, полученных посредством применения операции «печать–сканирование», необходимо провести двухэтапную обработку отсканированного изображения. На первом этапе осуществляется горизонтальное сканирование, определение положения границ линий текста и вертикальное расширение символов внутри отдельных строк. На втором этапе определяются границы слов и осуществляется удаление всех изолированных символов, ширина которых меньше установленного значения.

В отличие от предыдущих методов маркирования текстовых данных указанный метод не требует наличия исходного изображения в процессе извлечения данных. При этом вносятся ограничения на внедрение ЦВЗ только в тексты, выровненные по ширине, а также по наличию дополнительных этапов, направленных на коррекцию расширения (сжа-

тия) интервалов между словами с целью недопущения перекрытия слов и определение положения строк, слов, а также их границ после сканирования изображения.

В работе [17] предложен подход к маркированию текстовых документов, основанный на изменении величины интервала между словами. Встраивание информации в текстовые данные реализовано следующим образом:

– исходная информация подвергается помехоустойчивому кодированию (в качестве помехоустойчивого кода могут выступать код Хэмминга (7,4) или БЧХ (15,7));

– преобразование полученной кодовой последовательности в 16-битную периодическую m-последовательность посредством регистра сдвига;

– формирование 16-битового кода расширенного спектра вида {–1,1};

– сканирование текстового документа с целью обнаружения интервалов между словами;

– определение величин пустого пространства между словами каждой строки текста;

– изменение интервала между словами за счет расширения (сжатия) пустого пространства после каждого слова в строке на величину Δ (в зависимости от символа кодовой последовательности: «–1» соответствует уменьшению интервала, «1» – увеличению).

Извлечение встроенной информации может быть реализовано как из электронного текстового документа, так и из соответствующей печатной копии. Алгоритм извлечения встроенных данных из электронного документа состоит из следующих этапов:

– обнаружение и измерение пустого пространства между двумя последовательно идущими словами внутри каждой строки текста;

– вычисление среднеарифметического значения пустого пространства между словами относительно предшествующего и последующего интервала между словами;

– определение положения смещения слова посредством сравнения среднеарифметического значения пустого пространства с измеренным;

– декодирование полученной последовательности.

Извлечение встроенной информации из напечатанных на бумаге текстовых документов требует наличия дополнительных этапов:

– сканирование напечатанного текстового документа с разрешением не менее 300 точек на дюйм;

– преобразование отсканированного текстового изображения в бинарное;

– выравнивание ориентации изображения по линии текста;

– извлечение строк текста из бинарного изображения посредством построения горизонтального профиля проекции;

– определение расстояния между словами текста из горизонтального профиля проекции;

– корректировка полученных значений величин интервалов между словами;

– объединение полученных значений в массив данных;

– выполнение алгоритма извлечения встроенных данных из электронного документа.

В ходе экспериментальной оценки разработанного подхода к маркированию текстовых данных были проведены эксперименты, направленные на определение извлекаемости и точности извлечения встроенных данных.

В разработанном подходе к маркированию емкость встраивания зависит от количества текстовой информации, а также от параметров помехоустойчивого кода и кода расширенного спектра. Предельно достижимая емкость встраивания для текста, набранного гарнитурой Times New Roman с кеглем 11 пт, двойным межстрочным интервалом на листе формата Letter, составляет 300 бит. В ходе экспериментальной оценки предложенного алгоритма используемая длина внедряемой последовательности составляет 32 бита, при этом длина ЦВЗ составляет 8 бит.

В ходе оценки извлекаемости встроенных данных произведено извлечение встроенной информации из неподписанного текстового документа. Точность извлечения составляет 98,8%, что свидетельствует о наличии одиночных ошибок в процессе извлечения. Количественная оценка точности извлечения из подписанных электронных и напечатанных на бумаге текстовых документов не представлена. Однако авторами приведена информации о наличии как одинарных, так и двойных ошибок после декодирования полученных значений. Данный факт свидетельствует о высоком проценте ошибок в процессе извлечения данных и требует использования помехоустойчивых кодов с большей исправляющей способностью, что в свою очередь приводит к снижению предельно достижимой емкости встраивания.

Перцептивная невидимость встроенных данных основана на оценках, приведенных в работах [11–15]. Полученные ранее результаты позволяют отнести разработанный алгоритм к маркированию текстовых данных к перцептивно невидимым. Оценка робастности разработанного подхода к внесению искажений и осуществлению преобразований не проводилась.

Обсуждение результатов

Результаты анализа исследований в области маркирования текстовых данных позволяют перейти к сравнительной оценке разработанного подхода с рассмотренными аналогами. Сравнительная оценка проводилась по двум направлениям: по параметрам встраивания и извлечения информации, представленной водяным знаком, и по робастности встроенных данных к осуществлению преобразований и внесению искажений.

В ходе первого направления оценки разработанного подхода с рассмотренными аналогами получены результаты емкости встраивания, перцептивной невидимости, извлекаемости, точности извлечения и вычислительной сложности, представленные в табл. 3.

Анализ полученных результатов позволяет сделать вывод о том, что разработанный подход к мар-

кированию текстовых данных превосходит рассмотренные аналоги по точности извлечения встроенных данных, не накладывая требований по наличию исходного изображения в процессе извлечения данных. Кроме того, предложенный подход характеризуется низкой вычислительной сложностью. При этом предельно достижимая емкость встраивания превышает базовый аналог и [15], однако уступает [13, 14, 16]. Данная особенность обусловлена тем, что встраивание водяного знака производится в текстовые данные, а не в изображения, содержащие текст, как в [13, 14, 16].

В ходе оценки робастности встроенных данных рассмотренных подходов с разработанным подходом к осуществлению преобразований и внесению искажений получены результаты, представленные в табл. 4.

Таблица 3

Встраивание и извлечение информации в текстовые данные

Параметр	Исследование					
	Разработанный подход	[11–13]	[14]	[15]	[16]	[17]
Границы перцептивной невидимости	± 0,15 от исходного межстрочного интервала	Аналогично [7–10]	–	+	$0{,}1 \le C \le 0{,}2$	На основе оценок [11–17]
Емкость встраивания (бит)	60	23	1000	10000	42	300
Точность извлечения (%)	≥ 95	≤ 78	Данные отсутствуют	Данные отсутствуют	*	*
Наличие исходного документа	–	+	+	+	–	–
Вычислительная сложность	Низкая	Низкая	Высокая	Высокая	Средняя	Высокая

* Недостаточно данных для оценки параметра.

Таблица 4

Робастность встроенных данных к внесению искажений и осуществлению преобразований

Параметр	Исследование					
	Разработанный подход	[11–13]	[14]	[15]	[16]	[17]
Преобразование формата текстового документа в изображение	+	+	–	–	+	Данные отсутствуют
Перекопирование	До 5 раз	До 10 раз	–	–	До 10 раз	Данные отсутствуют
Поворот изображения (градусы)	На любой угол	≤ ±3	–	–	≤ ±5	Данные отсутствуют
Масштабирование (множитель)	±1,5	±0,04	0,5	0,5	–	Данные отсутствуют
Внесение шума	+	–	+	+	–	Данные отсутствуют
Сжатие изображения	+	–	JPEG (≤10%)	JPEG (≤10%)	–	Данные отсутствуют
Фильтрация	Медианная, гауссовская, модовая	Гауссовская	–	–	ФНЧ	Данные отсутствуют
Обрезка	≤ 50	–	≤ 75	≤ 75	–	Данные отсутствуют

Анализ полученных результатов позволяет сделать вывод о наличии инвариантности встроенных данных к рассмотренным преобразованиям. Кроме того, разработанный подход превосходит рассмотренные аналоги по большинству из оцениваемых параметров за исключением перекопирования [11–13, 16] и обрезки изображения [14, 15].

Полученные результаты позволяют отнести разработанный подход к маркированию текстовых данных к перцептивно невидимым подходам, отличающимся наличием инвариантности встроенных данных к основным операциям обработки изображений, а также к преобразованию формата текстовых данных в изображение, содержащее текст, посредством применения операции «печать–сканирование».

Заключение

Проведенный анализ параметров встраивания и извлечения разработанного подхода к маркированию текстовых данных с существующими аналогами позволяет сделать вывод о возможности применения данного подхода для защиты текстовой информации от утечки и защиты авторских прав владельцев данных. Кроме того, робастность встроенного водяного знака к преобразованию формата текстовых данных позволяет осуществлять идентификацию не только

электронных, но и печатных копий исходного текстового документа.

В то же время наличие ошибок в процессе извлечения встроенных данных и небольшая емкость встраивания накладывают ограничения на размер встраиваемой информации, представленной РВЗ. Ввиду этого снижение количества ошибок извлечения и повышение емкости встраивания являются направлением дальнейших исследований.

Литература

1. Через бумажные документы случается каждая девятая утечка конфиденциальных данных // Аналитический центр InfoWatch, 2019 [Электронный ресурс]. – Режим доступа: https://www.infowatch.ru/analytics/digest/15511 (дата обращения: 04.06.2019).

2. Alhindi H. Data Loss Prevention using document semantic signature / H. Alhindi, I. Traore, I. Woungang // International Conference on Wireless Intelligent and Distributed Environment for Communication. – 2018. – P. 75–99.

3. Wang C.-W. Data Loss Prevention system based on Big Data // 2nd International Conference on Artificial Intelligence: Techniques and Applications (AITA 2017). – 2017. – P. 292–298.

4. Stokes S. Digital copyright: law and practice. – Bloomsbury Publishing, 2019. – 297 p.

5. Eid A. A Tamper proofing text watermarking shift algorithm for copyright protection / A. Eid, A. Emran, A. Yahya // International Journal of Hybrid Information Technology. – 2018 – Vol. 11, No. 3. – P. 13–22.

6. Digital Watermarking Technique for Text Document Protection Using Data Mining Analysis / U. Khadam, M.M. Iqbal, M. Azam, S. Khalid // IEEE Access. – 2019. – Vol. 7. – P. 64955–64965.

7. Козачок А.В. Робастный водяной знак как способ защиты текстовых данных от утечки / А.В. Козачок, С.А. Копылов, М.В. Бочков // Защита информации. INSIDE (Санкт-Петербург). – 2018. – Т. 82, № 4. – С. 26–33.

8. Подход к извлечению робастного водяного знака из изображений, содержащих текст / А.В. Козачок, С.А. Копылов, Р.В. Мещеряков, О.О. Евсютин // Труды СПИИРАН (Москва). – 2018. – Т. 5(60). – С. 128–155.

9. Козачок А.В., Копылов С.А. Подход к внедрению робастного водяного знака в текстовые данные [Электронный ресурс]. – Режим доступа: http://www.ruscrypto.ru/resource/archive/rc2018/files/11_Kozachok_Kopylov.pdf, свободный (дата обращения: 05.02.2019).

10. Kozachok A.V. Text marking approach for data leakage prevention / A.V. Kozachok, S.A. Kopylov, A.A. Shelupanov, O.O. Evsutin // Journal of Computer Virology and Hacking Techniques. – 2019. DOI: 10.1007/s11416-019-00336-9 [Электронный ресурс]. – Режим доступа: https://link.springer.com/article/10.1007/s11416-019-00336-9 (дата обращения: 26.06.2019)

11. Marking text features of document images to deter illicit dissemination / J.T. Brassil, S. Low, N.F. Maxemchuk, L. O'Gorman // Proceedings of the 12th IARP International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. no. 94CH3440-5). – 1994. – Vol. 2. – P. 315–319.

12. Electronic marking and identification techniques to discourage document copying / J.T. Brassil, S. Low, N.F. Maxemchuk, L. O'Gorman // IEEE Journal on Selected Areas in Communications. – 1995. – Vol. 13, No. 8. – P. 1495–1504.

13. Brassil J.T. Copyright protection for the electronic distribution of text documents / J.T. Brassil, S. Low, N.F. Maxemchuk // Proceedings of the IEEE. – 1999. – Vol. 87, No. 7. – P. 1181–1196.

14. Secure spread spectrum watermarking for multimedia / I.J. Cox, J. Kilian, F.T. Leighton, T. Shamoon // IEEE transactions on image processing. – 1997. – Vol. 6, No. 12. – P. 1673–1687.

15. Marking and detection of text documents using transform-domain techniques / Y. Liu, J. Man, E. Wong, S.H. Low // Security and Watermarking of Multimedia Contents. International Society for Optics, Photonics. – 1999. – Vol. 3657. – P. 317–328.

16. Huang D. Interword distance changes represented by sine waves for watermarking text images / D. Huang, H. Yan // IEEE Transactions on Circuits and Systems for Video Technology. – 2001. – Vol. 11, No. 12. – P. 1237–1245.

17. Alattar A.M. Watermarking electronic text documents containing justified paragraphs and irregu-lar line spacing / A.M. Alattar, O.M. Alattar // Security, Steganography, and Watermarking of Multimedia Contents VI. T. 5306. International Society for Optics, Photonics. – 2004. – P. 685–696.

Козачок Александр Васильевич
Канд. техн. наук, сотрудник Академии
Федеральной службы охраны Российской Федерации
Приборостроительная ул., д. 35, г. Орел, Россия, 302034
Тел.: +7 (486-2) 54-99-33
Эл. почта: a.kozachok@academ.msk.rsnet.ru

Копылов Сергей Александрович
Сотрудник Академии
Федеральной службы охраны Российской Федерации
Приборостроительная ул., д. 35, г. Орел, Россия, 302034
Тел.: +7 (486-2) 54-99-33
Эл. почта: gremlin.kop@mail.ru

Kozachok A.V., Kopylov S.A.

The approach to text electronic documents marking and its comparison with existing analogues

This article presents the experimental evaluation results of the developed approach to the text documents marking. The features of the watermark being formed are considered. A review of existing research in the field of text data marking was conducted, the advantages and drawbacks of the considered approaches are determined. The results of a comparative analysis of the embedding and extraction parameters of the developed approach with the considered analogues are presented. The directions of further research aimed at increasing the embedding capacity and accuracy of data extraction were determined.

Keywords: information security, text data marking, text steganography.
doi: 10.21293/1818-0442-2019-22-2-52-60

References

1. InfoWatch. Every ninth leak of confidential data happens through paper documents. 2019. (In Russ.). Available at: https://www.infowatch.ru/analytics/digest/15511 (Accessed: June 4, 2019).

2. Alhindi H., Traore I., Woungang I. Data Loss Prevention using document semantic signature. International Conference on Wireless Intelligent and Distributed Environment for Communication, 2018, pp. 75–99.

3. Wang C.-W. Data Loss Prevention system based on Big Data. 2nd International Conference on Artificial Intelligence: Techniques and Applications (AITA 2017), 2017, pp. 292–298.

4. Stokes S. Digital copyright: law and practice. *Bloomsbury Publishing*, 2019, 297 p.

5. Eid A., Emran A., Yahya A. Tamper proofing text watermarking shift algorithm for copyright protection. *International Journal of Hybrid Information Technology,* 2018, vol. 11, no. 3, pp. 13–22.

6. Khadam U., Iqbal M.M., Azam M., Khalid S. Digital Watermarking Technique for Text Document Protection Using Data Mining Analysis. *IEEE Access,* 2019, vol. 7, pp. 64955–64965.

7. Kozachok A.V, Kopylov S.A., Bochkov M.V. Robust watermark as technique to text data leakage prevention. *Iinformation security INSIDE*, 2018, vol. 82, no. 4, pp. 26–33. (In Russ.).

8. 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. *SPIIRAS Proceedings*, 2018, vol. 60, no. 5, pp. 128–155 (in Russ.).

9. Kozachok A.V, Kopylov S.A. The embedding approach to robust watermarking in text data. *RusCrypto-2018*, 2018. (In Russ.). Available at: http://www.ruscrypto.ru/resource/archive/rc2018/files/11_Kozachok_Kopylov.pdf (Accessed: February 5, 2019).

10. Kozachok A.V., Kopylov S.A., Shelupanov A.A., Evsutin O.O. Text marking approach for data leakage prevention. *Journal of Computer Virology and Hacking Techniques*, 2019. DOI: 10.1007/s11416-019-00336-9 Available at: https://link.springer.com/article/10.1007/s11416-019-00336-9. (Accessed: June 26, 2019).

11. Brassil J.T., Low S., Maxemchuk N.F., O'Gorman L. Marking text features of document images to deter illicit dissemination. Proceedings of the 12th IARP International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5), 1994, vol. 2, pp. 315–319.

12. Brassil J.T., Low S., Maxemchuk N.F., O'Gorman L. Electronic marking and identification techniques to discourage document copying. *IEEE Journal on Selected Areas in Communications*, 1995, vol. 13, no. 8, pp. 1495–1504.

13. Brassil J.T., Low S., Maxemchuk N.F. Copyright protection for the electronic distribution of text documents. *Proceedings of the IEEE*, 1999, vol. 87, no. 7, pp. 1181–1196.

14. Cox I.J., Kilian J., Leighton F.T., Shamoon T. Secure spread spectrum watermarking for multimedia. *IEEE transactions on image processing,* 1997, vol. 6, no. 12, pp. 1673–1687.

15. Liu Y., Man J., Wong E., Low S. Marking and detection of text documents using transform-domain techniques. *Security and Watermarking of Multimedia Contents. International Society for Optics, Photonics*, 1999, vol. 3657, pp. 317–328.

16. Huang D., Yan H. Interword distance changes represented by sine waves for watermarking text images. *IEEE Transactions on Circuits and Systems for Video Technology*, 2001, vol. 11, no. 12, pp. 1237–1245.

17. Alattar A.M., Alattar O.M. Watermarking electronic text documents containing justified paragraphs and irregular line spacing. *Security, Steganography, and Watermarking of Multimedia Contents VI. T. 5306. International Society for Optics, Photonics*, 2004, pp. 685–696.

Alexander V. Kozachok
Candidate of Engineering,
Employee Academy of the Federal Guard Service
35, Priborostroitelnaya st., Orel, Russia, 302034
Phone: +7 (486-2) 54-99-33
Email: a.kozachok@academ.msk.rsnet.ru

Sergey A. Kopylov
Employee Academy of the Federal Guard Service
35, Priborostroitelnaya st., Orel, Russia, 302034
Phone: +7 (486-2) 54-99-33
Email: gremlin.kop@mail.ru