DOI: 10.15514/ISPRAS-2021-33(5)-4

Методы маркирования текстовых документов при печати посредством вертикального сдвига и изменения яркости фрагментов слов

[1] *Д.О. Обыденков, ORCID: 0000-0002-9296-6333 <obydenkov@ispras.ru>*
[1] *А.Е. Фролов, ORCID: 0000-0001-7616-2354 <aefrolov@ispras.ru>*
[1] *Ю.В. Маркин, ORCID: 0000-0003-1145-5118 <ustas@ispras.ru>*
[1] *С.А. Фомин, ORCID: 0000-0002-1151-2189 <fomin@ispras.ru>*
[2] *Б.В. Кондратьев, ORCID: 0000-0001-6348-117X <gae@mil.ru>*

[1] *Институт системного программирования им. В.П. Иванникова РАН, 109004, Россия, г. Москва, ул. Солженицына, д. 25*
[2] *Министерство обороны Российской Федерации, 119160, г. Москва, ул. Знаменка, д. 19.*

Аннотация. В статье представлены результаты разработки методов маркирования текстовых документов, представленных как растровое изображение. Важной особенностью алгоритмов является возможность обратного преобразования документа, что позволяет заменять метку на маркированном документе. Разработка относится к структурным алгоритмам маркирования на основе вертикальных сдвигов слов и изменения яркости отдельных фрагментов слов. В работе используются инструменты сегментирования для получения разметки текста в документе, БЧХ-коды для коррекции ошибок, метод максимизации правдоподобия для извлечения метки, нейронная сеть для восстановления искаженных слов. Тестирование показало практическую применимость разработанных алгоритмов маркирования при печати и сканировании текстовых документов.

Ключевые слова: маркирование документов; стеганография; обработка изображений; коррекция ошибок; алгоритмы машинного обучения; защита от утечек информации; слепое извлечение цифрового водяного знака

Для цитирования: Обыденков Д.О., Фролов А.Е., Маркин Ю.В., Фомин С.А., Кондратьев Б.В. Методы маркирования текстовых документов при печати посредством вертикального сдвига и изменения яркости фрагментов слов. Труды ИСП РАН, том 33, вып. 5, 2021 г., стр. 65-82. DOI: 10.15514/ISPRAS–2021–33(5)–4

Printed text documents watermarking based on vertical word shift and word fragments brightness changing

[1] *D.O. Obydenkov, ORCID: 0000-0002-9296-6333 <obydenkov@ispras.ru>*
[1] *A.E. Frolov, ORCID: 0000-0001-7616-2354 <aefrolov@ispras.ru>*
[1] *Y.V. Markin, ORCID: 0000-0003-1145-5118 <ustas@ispras.ru>*
[1] *S.A. Fomin, ORCID: 0000-0002-1151-2189 <fomin@ispras.ru>*
[2] *B.V. Kondrat'ev, ORCID: 0000-0001-6348-117X <gae@mil.ru>*

[1] *Ivannikov Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia*
[2] *Ministry of Defence of the Russian Federation, 19, Znamenka Str., Moscow, 119160, Russia*

Abstract. This paper describes the results of the development of methods for marking text documents represented as a raster image. An important feature of the algorithms is the possibility wipe current document mark and embed another one. The study refers to structural marking algorithms based on vertical word shifts and brightness changes of certain areas of the words. Segmentation tools are used to obtain document layout, BCH codes for error correction, a likelihood maximization method for label extraction, and a neural network for perturbed words recovery. Testing has proved the practical applicability of the algorithms with printing and scanning.

Keywords: data leakage prevention; documents watermarking; print-scan robust watermarking; blind watermarking method; stenography; image processing; error detection and correction; neural networks

For citation: Obydenkov D. O., Frolov A.E., Markin Y.V., Fomin S.A., Kondrat'ev B.V. Printed text documents watermarking based on vertical word shift and word fragments brightness changing**.** Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 5, 2021, pp. 65-82 (in Russian). DOI: 10.15514/ISPRAS–2021–33(5)–4

1. Введение

Многие частные компании и государственные учреждения несут финансовые и репутационные потери в связи с утечками конфиденциальных данных. Утечки можно разделить на две категории:

- *умышленные* – преднамеренные утечки, реализуемые с целью получения личной выгоды, промышленного шпионажа и прочего;
- *случайные* – прямо или косвенно произошедшие по неосторожности или неосведомленности сотрудников.

Согласно статистике [1] около половины утечек относится к первой категории. Для борьбы с неправомерным распространением конфиденциальной информации существуют различные программные системы, которые также можно разделить на два типа:

- *Системы предотвращения утечек* – системы, работающие превентивно и призванные не допустить утечку информации. Такие системы анализируют потоки данных, пересекающих границы внутренних сетей организации, потенциально способные содержать конфиденциальную информацию. Программные системы отслеживают различную «подозрительную активность» с использованием различных техник: анализ сетевого трафика, контроль чтения и записи информации на компьютерах сотрудников, отслеживание изменений сетевых настроек и т. д.;
- *Системы расследование инцидентов утечек* – системы, предназначенные для локализации и выяснения причин уже произошедшей утечки данных. Чаще всего подобные системы специальным образом помечают распространяемые данные, что позволяет по полученным в ходе расследования образцам определить источник утечки. Также схожие принципы могут применяться в системах защиты авторских прав.

Большинство систем предотвращения утечек данных, представленных на рынке, нацелены на работу исключительно с цифровыми каналами утечки информации: сеть, электронная почта, мессенджеры и прочие. Действительно, по ряду причин большинство утечек осуществляется именно по этим каналам. Однако методы противодействия утечкам информации на бумажных носителях по-прежнему актуальны – согласно отчету компании InfoWatch за 2020 год до 10% утечек приходится именно на этот канал [1]. Более того, в некоторых организациях доля утечек информации на бумажных носителях может быть выше в силу специфики рабочих процессов, как например, изолированная внутренняя сеть или запрет на использование личных мобильных смартфонов.

Текущая работа выполнялась в ходе проекта по созданию системы комплексной защиты от утечек информации. В данной статье рассматриваются методы ***внедрения*** метки в изображение выводимого на печать документа. *Меткой* в данном контексте является последовательность бит определенной длины, с помощью которой можно идентифицировать компьютер, с которого документ был отправлен на печать. Встраивание метки происходит путем малозаметного изменения внешнего вида текста документа. При расследовании инцидентов утечек выполняется ***извлечение*** метки из изображения маркированного (или *помеченного*) документа, полученного путем сканирования или фотографирования бумажного документа. При использовании бумажных носителей в системах документооборота зачастую возникает необходимость повторной печати отсканированных документов, потенциально уже маркированных, поэтому требуется обеспечить возможность замены метки на документах с уже встроенной меткой. Процесс замены метки можно разбить на две операции: ***стирание*** старой метки и внедрение новой метки. На Рис. 1 представлена схема маркирования документов, интегрированная в подсистему печати ОС [2], цветом выделены реализуемые алгоритмами маркирования операции.

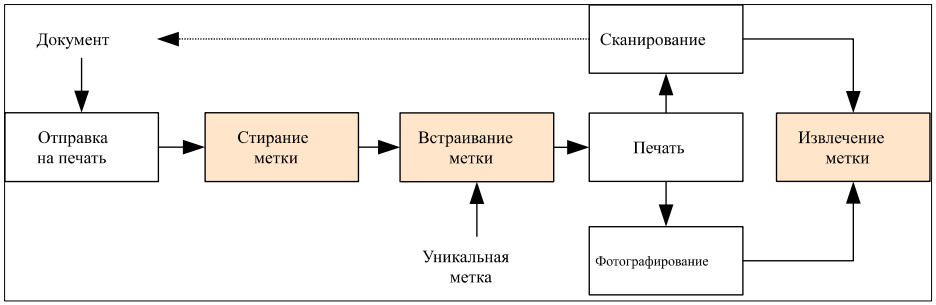


Рис. 1. Схема маркирования документов
Fig. 1. Document marking scheme

2. Связанные работы

Задача встраивания цифровых меток в изображения достаточно хорошо изучена и описана в ряде публикаций. Большинство авторов выделяют методы встраивания цифровых меток в область преобразования (*transform domain*) или в пространственная область (*spatial domain*).

К первой группе относятся методы встраивания цифровых меток в изображения, полученные путем некоторого преобразования – в частности, посредством дискретного косинусного преобразования (DCT) [3], дискретного преобразования Фурье (DFT) [4], дискретного вейвлет-преобразования (DWT). В данной статье эти методы рассматриваться не будут, поскольку их применение к изображениям текстовых документов (визуально близких к бинарным изображениям в отличие от цветных графических изображений) вносит существенные искажения.

Вторая группа включает методы маркирования, основанные на модификации оригинального изображения – в частности, внедрение некоторого периодического шаблона для графических изображений [5], сдвиг слов, изменение межстрочного интервала, добавление ошибок (орфографических, пунктуационных) для текстовых документов. Дальнейшая классификация методов маркирования пространственного домена выделяет лингвистические и структурные методы [6][7]:

- **Лингвистические методы** — изменяют семантические или синтаксические свойства текстового содержимого документа;
- **Структурные методы** — изменяют параметры визуального представления текстового слоя документа, но не меняют смысл текста.

2.1. Лингвистические методы

Лингвистические методы маркирования документов подразделяются на категории:

- *Семантические* — включают в себя различные методы сокрытия информации в тексте посредством изменения таких атрибутов текста как правописание, использованием аббревиатур и акронимов, замена на синонимы и другие.
- *Синтаксические* — сокрытие информации без значительного изменения смысла текста. В значительной степени такие методы маркирования полагаются на синтаксические свойства языка или использует присущие ему особенности. Например, замена определенных букв некоторого алфавита на визуально схожие буквы другого алфавита.

2.2. Структурные методы

Структурные методы маркирования документов подразделяются на следующие категории.

- *Непечатаемые символы*

  Различные непечатаемые символы широко используются при маркировании документов. Существуют работы, где такие символы используются как для замены, так и для вставки содержимого в исходном тексте. Пробельные символы хорошо подходят для встраивания информации в текстовый слой, поскольку крайне сложно визуально отличить различные вариации пробела (в Unicode-кодировке имеется почти два десятка пробельных символов). Встречаются различные вариации алгоритмов, когда информация встраивается между словами, предложениями, абзацами, пустыми строками и иные [8]-[10].

  Также существуют алгоритмы, использующие для кодирования символы нулевой ширины, например, пробел нулевой длины или символы, выполняющие сервисные функции в лигатурных шрифтах. Всего в Unicode существует 10 непечатаемых символов. Кодирующие метку символы могут размещаться на самых различных позициях, например, после пунктуационных символов, между пустыми линиями и параграфами [11][12].

- *Смещение текстовых элементов*

  Одна из первых работ по маркированию документов на основе вертикального смещения текстовых линий появилась еще в 1994 году [13]. Смещение текстовой линии вверх или вниз кодировало единичный бит [15]. Позднее появились работы, в которых для кодирования применялось горизонтальное смещение слов [14]: слова объединяются в группы по три и слово, стоящее посередине, смещается вправо или влево для кодирования одного бита метки [16][17].

- *Свойства форматирования текста*

Большим разнообразием обладает группа методов, использующих для кодирования различные свойства форматирования текста: размер, цвет, особенности начертания шрифта и другие свойства. В частности, разработано множество методов кодирования метки на основе искажения начертания глифов. В ранних работах вносимые искажения довольно грубые и визуально заметны. В последующих работах изменения менее заметны, поскольку модифицируются лишь контуры отдельных глифов [18]. Метод на основе искажения шрифта использует нейронную сеть для считывания вариации конкретной буквы и визуально различим только при близком рассмотрении и прямом сравнении с оригиналом [19].

Также существует группа методов, специализирующаяся на кодировании с использованием особенностей арабского языка. Метка кодируется посредством сдвигов точек в определенных словах, используемых при начертании букв арабского алфавита, а также удлинения вертикальных черт в других буквах, что не меняет смысла написанного [20].

2.3. Существующие решения

В настоящий момент на российском рынке существует несколько продуктов, которые реализуют возможность внедрения уникальных меток в выводимые на печать документы: *TraceDoc* [21] и *EveryTag* [22]. Информация об используемых методах маркирования представляет коммерческую тайну и отсутствует в открытом доступе, однако описание продуктов на сайтах производителя позволяет сделать вывод, что для извлечения метки из маркированного документа системам требуется оригинальная, немаркированная версия документа. Такой подход имеет свои преимущества – возможность сравнивать маркированный документ с оригиналом потенциально повышает точность извлечения, а также снижает заметность метки на документе. В то же время необходимость содержания и поддержки базы данных оригиналов и всех уникальных копий документов является значительным недостатком данного подхода, так как такая база становится единой точкой отказа, а также централизованный характер такой базы может быть несовместим со структурой некоторых организаций.

2.4. Резюме

Среди рассмотренных методов маркирования был выбран структурный подход как наиболее подходящий для решения поставленной задачи – маркирование текстовых документов. Лингвистические методы не подходят, поскольку изменяют содержимое документа, что в имеющейся постановке недопустимо. Методы встраивания цифровых меток в область преобразования также не подходят, так как существенно ухудшают визуальное представление документа. Подходы к маркированию на основе изменения свойств отдельных глифов чрезвычайно требовательны к качеству изображения при извлечении, что сильно ограничивает их возможное практическое применение. Методы встраивания цифровых меток, использующие особенности определенных форматов, ограничивают применение сравнительно небольшим числом форматов документов. Также при встраивании используется структура формата, предоставляющая информацию о текстовом слое, форматировании и других атрибутах документа, тогда как при извлечении или повторном встраивании сканированный или фотографированный документ представлен как растровое изображение.

Разработка метода кодирования требует поиска компромисса между противоречивыми критериями:

- **Незаметность** – стойкость метода к обнаружению метки в маркированном документе;

- **Надежность** – стойкость метода к различным искажениям, возникающим при печати/сканировании/фотографировании маркированного документа;
- **Емкость** – количество бит, которые можно встроить в маркируемый документ.

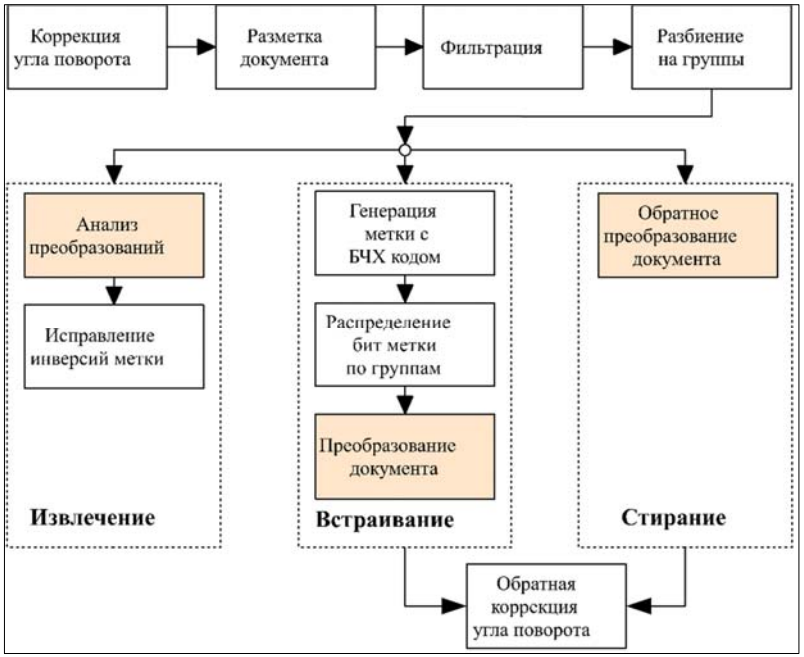


Рис. 2. Этапы маркирования документа для различных операций (цветом выделены этапы, реализация которых зависит от алгоритма)
Fig. 2. Document marking process pipeline for different operations (algorithm specific stages colored)

3. Описание алгоритмов маркирования

В рамках работы над проектом было разработано и реализовано два метода маркирования текстовых документов. В основе первого метода — незначительное **смещение** слова по вертикали. Второй метод изменяет яркость начертания шрифта вдоль линии, проходящей посередине слова – визуально данный эффект напоминает **перечеркивание** слова.

Алгоритмы рассчитаны на работу с растровыми изображениями текстовых документов, разрешение которых превышает 150 DPI. Каждая страница многостраничного текстового документа маркируется по отдельности. На каждой странице документа должно присутствовать не менее 12 строк текста, а в каждой строке – не менее пяти слов для маркирования посредством первого алгоритма и не менее трех слов для маркирования с помощью второго.

3.1. Общая схема работы

Несмотря на то, что механизмы работы разработанных алгоритмов кодирования различны, значительная часть логики обработки документов совпадает. На Рис. 2 представлена последовательность выполнения этапов для различных операций маркирования документа. Общие этапы обработки документа, представленные на схеме, включают:

- *коррекция угла поворота*: корректировка осуществляется в небольшом диапазоне углов и не исправляет ошибки ориентации документа;
- *разметка документа*: выделение на изображении документа слов, текстовых линий, текстовых блоков;
- *фильтрация*: составление набора элементов разметки (слов или текстовых линий), которые не будут использоваться при кодировании;
- *разбиение на группы*: объединение слов в текстовых линиях в группы, кодирующих один бит метки;
- *обратная коррекция угла поворота*: восстановление исходного угла поворота документа.

3.1.1. Встраивание метки

Входные данные: растровое изображение документа, 32-битная метка;

Выходные данные: растровое изображение документа с внедренной меткой.

Краткое описание этапов обработки во время встраивания метки:

- *Генерация метки с БЧХ кодом*: преобразование входной метки в последовательность битов, добавление блока коррекции ошибок на основе БЧХ-кода;
- *Распределение бит метки по группам*: вычисление соответствий между битами метки и группами, которые их кодируют;
- *Преобразование документа* в соответствии с выбранным методом представления метки: перечеркивания (раздел 4) или смещения (раздел 5).

3.1.2. Извлечение метки

Входные данные: растровое изображение документа с внедренной меткой;

Выходные данные: 32-битная метка.

Краткое описание этапов обработки во время извлечения метки:

- *Анализ преобразований*: извлечение закодированных бит из каждой группы, объединение в битовую последовательность. Реализация данного этапа зависит от выбранного метода: перечеркивания (раздел 4) или смещения (раздел 5);
- *Исправление инверсий метки*: разделение извлеченной битовой последовательности на метку и БЧХ-коды коррекции. Обнаружение инверсий бит, если количество ошибок не слишком велико.

3.1.3. Стирание метки

Входные данные: растровое изображение документа с внедренной меткой;

Выходные данные: растровое изображение документа со стертой меткой.

Краткое описание этапов обработки во время стирания метки:

- *Обратное преобразование документа* осуществляется по группам. Реализация данного этапа зависит от выбранного метода: перечеркивания (раздел 4) или смещения (раздел 5)

3.2. Разметка документа

Разработанные алгоритмы маркирования оперируют над словами и текстовыми линиями, поэтому один из этапов работы алгоритмов – получение разметки текста. *Результат выполнения разметки текста* — это информация о минимальных ограничивающих слова прямоугольниках, базовых линиях и взаимном расположении элементов на изображении текстового документа. *Минимальный ограничивающий прямоугольник (МОП)* —

прямоугольник минимальной площади, внутри которого слово или текстовая линия находится целиком. *Базовая линия* — воображаемая линия, проходящая по нижнему краю слова без учета нижних выносных элементов (пример на Рис. 3).

Для получения разметки текста используются инструменты сегментации. В данный момент поддерживается два инструмента:

- **Tesseract OCR** — это свободно распространяемая программа для распознавания текста [23];
- **DocParser** — решение, разработанное партнерами ИСП РАН на основе нейронной сети архитектуры UNet [24].

При работе с инструментами сегментации был выявлен ряд проблем. Главная проблема – несовпадение разметки оригинально документа и разметки сканированного изображения того же документа. Такое несовпадение может сильно повлиять на корректность внедрения и точность извлечения метки. В ходе работы были выделены следующие типы ошибок:

- разбиение одного слова на два;
- слияние двух слов в одно;
- выделение различных дефектов (шум, дефекты сканирования) как отдельных текстовых элементов;
- некорректная работа с документами, содержащими таблицы;
- некорректная работа с документами, маркированными *методом перечеркивания*.

3.2. Механизмы обработки ошибок

Вышеизложенные проблемы инструментов сегментации текста влияют на корректность извлечения метки. Для повышения точности алгоритмов было разработано несколько техник, которые позволяют значительно уменьшить влияние ошибок разметки на корректность извлечения метки. Среди них:

- **Группировка слов в текстовых линиях**. Суть данного метода заключается в том, что один бит метки кодируется не одним словом, а группой из нескольких слов. Каждая строка документа делится на N (по умолчанию $N = 3$) равных по ширине частей, каждое слово в текстовой линии попадет в определенную группу в зависимости от своего положения. Группа может включать несколько слов или быть пустой. Данный метод позволяет частично нивелировать ошибки разметки, связанные с некорректным определением границ слов.
- **Коды коррекции ошибок**. Метка, которая внедряется в документ, при необходимости может быть дополнена блоком обнаружения и исправления битовых инверсий. Для коррекции ошибок используется БЧХ-код [25], который в теории позволяет корректировать произвольное количество инверсий, однако увеличение количества обнаруживаемых ошибок требует увеличения длины блока коррекции. По умолчанию корректируется 3 ошибки – длина блока с кодами коррекции при этом составляет 18 бит.
- **Фильтрация элементов**. Некоторые слова или текстовые линии могут не подходить для кодирования информации, ввиду того что они являются источником ошибок при разметке или из-за специфики конкретного алгоритма кодирования. При внедрении и извлечении метки такие элементы пропускаются и не используются для кодирования. Используются следующие критерии: ширина слов, количество слов в текстовой линии, нахождение слова в таблице и т.д.
- **Детектирование таблиц**. Инструменты сегментации плохо справляются с разметкой документов, в которых присутствуют таблицы. Для решения этой проблемы все слова,

которые находятся внутри границ таблиц, отфильтровываются. Разработанный алгоритм детектирования таблиц находит внешние контуры всех таблиц в документе и помечает слова, находящиеся внутри этих контуров. Поиск внешних контуров осуществляется на основе топологического анализа бинаризованных изображений [26].

3.3. Позиционирование метки в документе

Если в документе присутствует больше текстовых линий, чем необходимо для однократного внедрения метки, выполняется поиск наиболее подходящей области встраивания метки. Область встраивания представляет собой набор текстовых линий, следующих друг за другом, ограниченный специальными строками-маркерами. При выборе позиции для встраивания метки учитываются следующие факторы (в порядке приоритета):

1. наличие в строке слов, не являющихся машинописным текстом[1] (меньше — лучше);
2. количество пустых групп (меньше — лучше);
3. «заполненность» текстовой линии, вычисляемая по формуле ниже (больше — лучше).

$$tl^i_{fat} = \frac{tl^i_{width}}{tl^W_{width}} \cdot \frac{\sum^{\overline{tl}^i} w^{i,j}_{area}}{tl^i_{area}}$$

- tl^i_{width} — ширина МОП i текстовой линии;
- tl^i_{area} — площадь МОП i текстовой линии;
- $w^{i,j}_{area}$ — площадь МОП j слова в i текстовой линии;
- $\overline{tl}^i$— количество слов в i текстовой линии;
- W — номер самой широкой текстовой линии.

Область встраивания метки – заданный диапазон строк – ограничивается посредством специальных строк-маркеров начала и конца метки. В качестве метода преобразования для строк-маркеров начала и конца метки используется алгоритм перечеркивания слов.

4. Алгоритм кодирования на основе перечеркивания слов

4.1. Внедрение метки

Алгоритм кодирования использует подход с изменением яркости глифов в областях пересечения горизонтальной линии, проходящей по длине слова между *базовой* и *медианной* линиями, а также имеющей толщину σ. Визуально данный эффект напоминает перечеркивание слова (рис. 3). Толщина линии σ вычисляется на основе средней высоты слов во всем документе. Перечеркнутое слово кодирует единичный бит, неперечеркнутое слово – нулевой (рис. 4).

Рис. 3. Расположение перечеркивающей линии
Fig. 3. Strikethrough line location

[1] Для определения типа содержимого слова используется нейронная сеть на основе архитектуры ResNet [27].

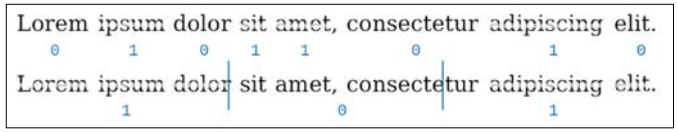


Рис. 4. Перечеркивание слов: обычное (выше), с группировкой (ниже)
Fig. 4. Words strikethrough: simple (upper), with grouping (below)

4.2. Извлечение метки

Инструменты получения разметки документа плохо работают с документами, в которые внедрена метка посредством перечеркивания слов. Для решения этой проблемы перед получением разметки проводится операция **восстановления документа**, которая основана на морфологической операции *эрозии*[2]. Для корректной работы алгоритма восстановления требуется знать значение толщины перечеркивающей линии в пикселях σ. На этапе внедрения метки данная величина динамически вычисляется, используя полученное в ходе разметки значение средней высоты строк. Для алгоритма восстановления был выбран подход, который аппроксимирует среднюю высоту строк без использования разметки документа.

После получение разметки алгоритм извлечения метки для каждого слова определяет, перечеркнуто оно или нет. Алгоритм детектирования перечеркивания в слове работает следующим образом:

1. Фрагмент изображения, содержащий слово, бинаризуется;
2. Проводится множество линий от левого края МОП слова к правому. Если нашлась хотя бы одна линия, не пересекающая черных пикселей, то слово перечеркнуто и кодирует 1, иначе – 0.

4.3. Стирание метки

На текущий момент существует два подхода к стиранию метки: морфологический и нейросетевой. *Морфологический* подход использует описанный в разделе 4.2 алгоритм восстановления документа. К восстановленному документу применяется операция *эрозии*, вычисляются области перечеркивания слов, после чего фрагменты слов из восстановленного документа накладываются на исходный документ. Качество работы алгоритма стирания метки зависит, в том числе от содержащихся в слове букв. Например, в кириллических буквах *а, в, е, н, ю, э, и, й, ж, з* высока вероятность появление дефектов.

Нейросетевой алгоритм стирания метки, использующий сверточную нейронную сеть UNet [24], дает меньшее количество дефектов (рис. 5). Архитектура UNet была выбрана среди прочих, поскольку хорошо справляется с задачей сегментации изображений. Полная сверточность нейросети позволяет в случае необходимости изменить размер входного изображения. Для обучения нейронной сети был собран набор данных, состоящий из 20427 изображений слов. На каждое изображение слова накладывалось перечеркивание с определенными параметрами и генерировалась соответствующая *маска перечеркивания*. Бинаризованное изображение перечеркнутого слова является входом нейронной сети, а ожидаемым выходом является *маска перечеркивания*.

[2] Операция свертки изображения с прямоугольным ядром. Визуально данная операция вызывает расширение темных областей на изображении.

Рис. 5. Сравнение методов стирания. Сверху слово с внедренной меткой, посередине - морфологический подход, снизу - нейросетевой подход
Fig. 5. Wipe methods comparison. On top word with label injected, on the middle - morphological approach, bottom - neural network approach

5. Алгоритм кодирования на основе вертикального смещения слов

Как упоминалось ранее, среди структурных методов кодирования распространены алгоритмы, использующие смещение текстовых линий и параграфов по вертикали или слов по горизонтали. Однако не встречается работ на основе вертикального смещения слов, поэтому была предпринята попытка реализовать подобный алгоритм.

5.1 Внедрение метки

В ходе работы был разработан алгоритм кодирования метки на основе вертикального смещения слов. Первое и последнее слово в строке обозначаются *опорными* и не смещаются, что позволяет задать нулевой уровень смещения и оценивать смещение остальных слов относительно *опорных*. Уровень слова определяется по его *базовой линии*. Данный метод использует следующую схему кодирования: если слово кодирует 1, то оно смещается вверх на β пикселей, если 0 — остается на месте. Значение β вычисляется на основе средней высоты слов во всем документе. При использовании группировки все слова, находящиеся в одной группе, кодируют один бит, а значит, имеют одно и то же смещение. Пример кодирования вертикальным смещением представлен на Рис. 6, величина β намеренно увеличена для наглядности.

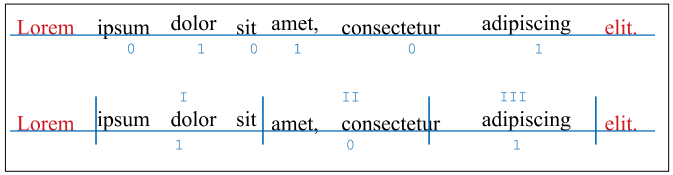


Рис. 6. Вертикальное смещение слов: обычное (выше), с группировкой (ниже)
Fig. 6. Vertical word-shift: simple (upper), with grouping (below)

При смещении все содержимое МОПа слова смещается вместе с ним. Фрагмент изображения, который при смещении перекрывается словом, также смещается и замещает освободившееся пространство.

5.2. Извлечение метки

Алгоритм извлечения метки работает построчно и определяет смещение для каждого слова в строке. При считывании закодированной последовательности используется смещение базовой линии слова относительно предыдущего. Наиболее вероятная последовательность смещений слов в строке вычисляется при помощи алгоритма Витерби. Информация о том, что первое и последнее слова в строке не смещены, позволяет отсечь множество вариантов и выбрать только один. После считывания последовательности из строки определяется соответствие между словом и закодированным битом. В силу искажений различного

характера и ошибок разметки возможна ситуация, когда словам в одной группе соответствуют разные биты, поэтому при выборе бита группы вычисляется взвешенная сумма для каждого значения. Вес для каждого слова вычисляется индивидуально на основе следующих факторов:

1. ширина слова (больше — лучше);
2. количество символов в слове (больше — лучше);
3. наличие символов с выносными элементами[3] (меньше — лучше).

5.3. Стирание метки

При стирании метки используется информация о смещениях в каждой строке, полученная при помощи алгоритма Витерби аналогично тому, как это делается при извлечении метки. После извлечения битовой последовательности из строки выполняется обратное смещение для каждого слова индивидуально, без учета распределения слов по группам.

Строки-маркеры начала и конца области встраивания метки восстанавливаются при помощи тех же алгоритмов, что используются при стирании в алгоритме маркирования методом перечеркивания слов.

6. Результаты тестирования

Разработанные в ходе исследования алгоритмы были реализованы и протестированы. Инструментов маркирования текстовых документов на основе структурных алгоритмов найдено не было. Для тестирования алгоритмов была собрана выборка из 40 документов различного содержимого и форматирования. Документы для данной выборки были загружены с сайта Министерство образования и науки РФ (URL: https://minobrnauki.gov.ru/). В каждый документ внедрялось 4 метки:

1) случайная метка (сгенерированная перед тестированием),
2) инверсия случайной метки,
3) метка, состоящая только из единиц,
4) метка, состоящая только из нулей.

Испытания проводились по трем сценариям (**П** – печать, **С** – сканирование, **Ф** – фотографирование):

- **ПС**. В документ внедряется метка, документ с меткой печатается, распечатанный документ сканируется, из скана документа извлекается метка;
- **ПСПС**. Из скана документа из первого сценария стирается метка, в документ со стертой меткой внедряется инверсия предыдущей метки, документ с новой меткой печатается, распечатанный документ сканируется, из скана документа извлекается метка. В данном сценарии использовались только случайная метка и ее инверсия;
- **ПФ**. В документ внедряется метка, документ с меткой печатается, распечатанный документ фотографируется, из фотографии документа извлекается метка. В данном сценарии применялись только случайная метка и ее инверсия.

Для оценки результатов используются следующие метрики:

- **Внедрено** – доля документов, в которые удалось встроить метку. Некоторые документы не могут быть использованы для встраивания метки определенным алгоритмом ввиду недостаточного количества текста;
- **Извлечено** – доля документов, из которых удалось извлечь метку;
- **Точность** — средняя доля одинаковых бит при внедрении и извлечении метки (без

[3] Наличие нижних выносных элементов повышает вероятность корректного вычисления базовой линии.

блока с кодами коррекции);

- **Полнота относительная** — доля извлеченных меток с точностью равной 1;
- **Полнота абсолютная** – доля извлеченных меток с точностью равно 1 среди всех документов.

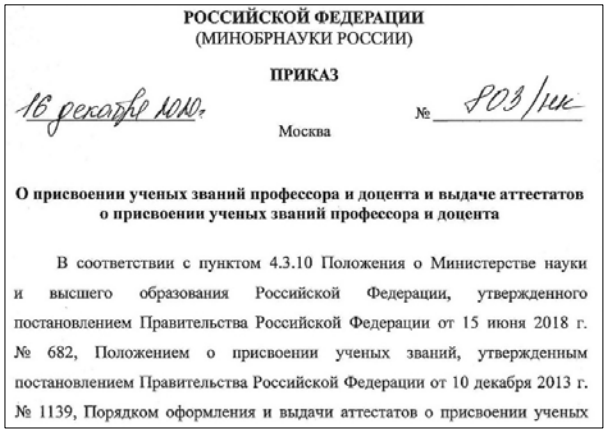
Алгоритм маркирования на основе вертикального смещения слов показывает лучшие значения метрик абсолютной полноты, что дает кумулятивную оценку точности и универсальности (табл. 1, 2, рис. 7-9)). Большее число успехов при внедрении метки первым алгоритмом обуславливается меньшим количеством слов в текстовой линии, необходимым для разбиения на группы. Снижение точности извлечения в сценарии *ПФ* по сравнению с другими сценариями объясняется значительными искажениями, в частности, аберрацией и дисторсией.

Табл. 1. Результаты тестирования алгоритмом на основе смещения слов
Table 1. Test results of the algorithm based on word shift

Тестовый сценарий	*Смещение*				
	Внедрено	*Извлечено*	*Точность*	*Полнота относительная*	***Полнота абсолютная***
ПС	88/160	74/160	0.980997	0.864865	0.4000
ПСПС	88/160	71/160	0.898768	0.521127	0.2312
ПФ	44/80	35/80	0.829464	0.285714	0.1249

Табл. 2. Результаты тестирования алгоритмом на основе перечеркивания слов
Table 2. Test results of the algorithm based on word strikethrough

Тестовый сценарий	*Перечеркивание*				
	Внедрено	*Извлечено*	*Точность*	*Полнота относительная*	***Полнота абсолютная***
ПС	148/160	147/160	0.966837	0.897959	0.8249
ПСПС	148/160	148/160	0.944046	0.871622	0.8062
ПФ	40/40	39/40	0.737981	0.282051	0.2749


РОССИЙСКОЙ ФЕДЕРАЦИИ
(МИНОБРНАУКИ РОССИИ)

ПРИКАЗ

16 декабря 2010 г. № 803/нк

Москва

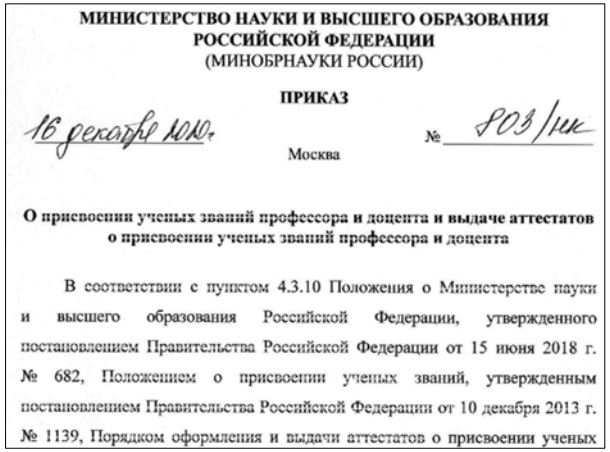
О присвоении ученых званий профессора и доцента и выдаче аттестатов о присвоении ученых званий профессора и доцента

В соответствии с пунктом 4.3.10 Положения о Министерстве науки и высшего образования Российской Федерации, утвержденного постановлением Правительства Российской Федерации от 15 июня 2018 г. № 682, Положением о присвоении ученых званий, утвержденным постановлением Правительства Российской Федерации от 10 декабря 2013 г. № 1139, Порядком оформления и выдачи аттестатов о присвоении ученых

Рис. 7. Фрагмент отсканированного документа без метки
Fig. 7. Fragment of scanned document without label

7. Заключение

В рамках исследования были разработаны алгоритмы маркирования выводимых на печать текстовых документов на основе вертикального сдвига и перечеркивания слов. Возможность извлечения метки без использования оригинального документа является преимуществом перед продуктами, которые уже присутствуют на рынке, а возможность стереть метку из маркированного документа является нововведением по сравнению с подходами, которые описаны в опубликованных ранее работах. В ходе работы было проведено тестирование программных реализаций алгоритмов на всех предполагаемых сценариях использования, включая стирание метки и ее повторное внедрение. Результаты тестирования показали практическую применимость разработанных методов маркирования.


МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
(МИНОБРНАУКИ РОССИИ)

ПРИКАЗ

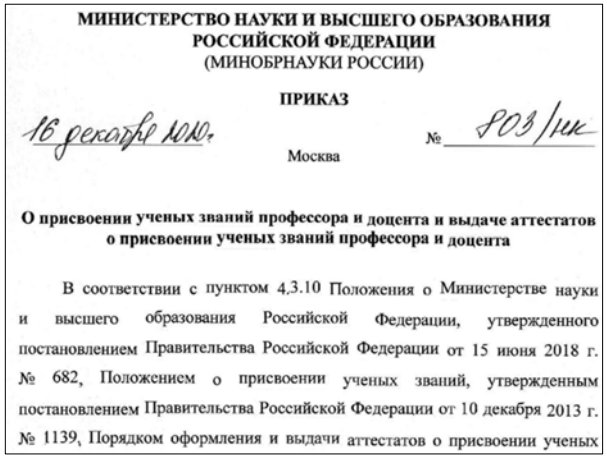
16 декабря 2010 г. № 803/нк

Москва

О присвоении ученых званий профессора и доцента и выдаче аттестатов о присвоении ученых званий профессора и доцента

В соответствии с пунктом 4.3.10 Положения о Министерстве науки и высшего образования Российской Федерации, утвержденного постановлением Правительства Российской Федерации от 15 июня 2018 г. № 682, Положением о присвоении ученых званий, утвержденным постановлением Правительства Российской Федерации от 10 декабря 2013 г. № 1139, Порядком оформления и выдачи аттестатов о присвоении ученых

Рис. 8. Фрагмент отсканированного документа с меткой на основе алгоритма перечеркивания слов
Fig. 8. Fragment of scanned document with label based on word strikethrough algorithm


МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
(МИНОБРНАУКИ РОССИИ)

ПРИКАЗ

16 декабря 2010 г. № 803/нк

Москва

О присвоении ученых званий профессора и доцента и выдаче аттестатов о присвоении ученых званий профессора и доцента

В соответствии с пунктом 4.3.10 Положения о Министерстве науки и высшего образования Российской Федерации, утвержденного постановлением Правительства Российской Федерации от 15 июня 2018 г. № 682, Положением о присвоении ученых званий, утвержденным постановлением Правительства Российской Федерации от 10 декабря 2013 г. № 1139, Порядком оформления и выдачи аттестатов о присвоении ученых

Рис. 9. Фрагмент отсканированного документа с меткой на основе алгоритма вертикального смещения слов
Fig. 9. Fragment of scanned document with label based on word shift algorithm

Пути дальнейшего развития включают следующие направления:

- повышение качества работы при извлечении метки из фотографии;
- уменьшение визуальной заметности метки;
- разработка методов эмуляции дефектов, которые могут вносить сканеры, что позволит значительно упростить тестирование;
- повышение стабильности работы инструментов сегментации текста на сканах и фотографиях.

Список литературы / References

[1] Утечки информации ограниченного доступа: отчет за 9 месяцев 2020 г. Экспертно-аналитический центр InfoWatch, 2020 г. / Restricted information leaks: report for 9 months of 2020. InfoWatch Analytical Center, 2020 (in Russian).

[2] Козлов С.В., Копылов С.А. и др. Реализация маркирования в подсистеме печати ОС семейства Windows на основе виртуального XPS-принтера. Труды ИСП РАН, том 32, вып. 5, 2020 г., стр. 95-110 / Kozlov S.V., Kopylov S.A. et al. Implementing watermarking based on a virtual XPS printer for Windows operating systems. rudy ISP RAN/Proc. ISP RAS, vol. 32, issue 5, 2020, pp. 95-110 (in Russian). DOI: 10.15514/ISPRAS–2020–32(5)–7.

[3] Dong P., Galatsanos N. P. Affine transformation resistant watermarking based on image normalization. In Proc. of the International Conference on Image Processing, 2002, pp. 489-492.

[4] Pramila A., Keskinarkaus A., Seppänen T. Multiple domain watermarking for print-scan and JPEG resilient data hiding. Lecture Notes in Computer Science, vol. 5041, 2007, pp. 279-293.

[5] Ahmed Q., Munib S., Mirza M. T., Khan A. Smart phone based online medicine authentication using print-cam robust watermarking. In Proc. of the 13th International Conference on Frontiers of Information Technology (FIT), 2015, pp. 222-227.

[6] Ahvanooey M.T., Li Q. et al Modern text hiding, text steganalysis, and applications: a comparative analysis. Entropy, vol. 21, no. 4, 2019, article 355.

[7] 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.

[8] Por L. Y., Wong K., Chee K. O. UniSpaCh: A text-based data hiding method using Unicode space characters. Journal of Systems and Software, vol. 85, no. 5, 2012, pp. 1075-1082.

[9] Bender W., Gruhl D. et al. Techniques for data hiding. IBM Systems Journal, vol. 35, issue 3.4, 1996, pp. 313-336.

[10] Leea I.S. Secret communication through web pages using special space codes in HTML files. International Journal of Applied Science and Engineering, vol. 6, no. 2, 2008, pp. 141-149.

[11] Ahvanooey M.T., Tabasi S.H. A new method for copyright protection in digital text documents by adding hidden unicode characters in persian/english texts. International Journal of Current Life Sciences, vol. 4, no. 8, 2014, pp. 4895-4900.

[12] Ahvanooei M.T., Tabasi S.H., Rahmani S. A novel approach for text watermarking in digital documents by zero-width interword distance changes. DAV International Journal of Science, vol. 4, no. 3, 2015, pp. 550-558.

[13] Low S.H., Maxemchuk N.F. et al. Document marking and identification using both line and word shifting. In Proc. of the INFOCOM'95, 1995, pp. 853-860.

[14] Brassil J. T., Low S. et al. Electronic marking and identification techniques to discourage document copying. IEEE Journal on Selected Areas in Communications, vol. 13, issue 8, 1995, pp. 1495-1504.

[15] Alattar A.M., Alattar O.M. Watermarking electronic text documents containing justified paragraphs and irregular line spacing. Proceedings of the SPIE, vol. 5306, Security, Steganography, and Watermarking of Multimedia Contents VI, 2004, pp. 685-695.

[16] Kim Y.W., Moon K.A., Oh I.S. A Text Watermarking Algorithm based on Word Classification and Inter-word Space Statistics. In Proc. of the Seventh International Conference on Document Analysis and Recognition, 2003, pp. 775-779.

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

[18] Tan L., Hu K. et al. Print-scan invariant text image watermarking for hardcopy document authentication. Multimedia Tools and Applications, vol. 78, no. 10, 2018, pp. 13189-13211.

[19] Xiao C., Zhang C., Zheng C. Fontcode: Embedding information in text documents using glyph perturbation. ACM Transactions on Graphics (TOG), vol. 37, no. 2, 2017, pp. 1-16.

[20] Gutub A., Fattani M. A novel Arabic text steganography method using letter points and extensions. In Proc. of the WASTET International Conference on Computer, Information and System Science and Engineering (ICCISSE), 2007, pp. 28-31.

[21] Secret Technologies – Trace Doc. Available at https://secretgroup.ru/trace-doc/, accessed 09.04.2021.

[22] EVERYTAG – Information Leaks Detection (ILD). Available at https://everytag.ru/, accessed 09.04.2021.

[23] Smith R. An overview of the Tesseract OCR engine. In Proc. of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), 2007, pp. 629-633.

[24] Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science, vol. 9351, 2015, pp. 234-241.

[25] Морелос-Сарагоса Р. Искусство помехоустойчивого кодирования. Техносфера, 2005, 320 стр. / Morelos-Zaragoza R. The Art of Error Correcting Coding. Willey, 2002, 238 p.

[26] Suzuki S. Topological structural analysis of digitized binary images by border following. Computer vision, graphics, and image processing, vol. 30, no. 1, 1985, pp. 32-46.

[27] He K., Zhang X. et al. Deep residual learning for image recognition. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

Информация об авторах / Information about authors

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

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

Александр Евгеньевич ФРОЛОВ – студент-магистр. Научные интересы: стеганография, методы анонимизанции сетевого трафика, компьютерные сети, методы машинного обучения.

Alexander Evgenevich FROLOV is a master student. Research interests: steganography, traffic anonymization methods, computer networks, machine learning.

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

Yury Vitalievich MARKIN, researcher, PhD. Research interests: information security, network traffic analysis, image processing, machine learning algorithms.

Станислав Александрович ФОМИН — ведущий программист. Область научных интересов: теория сложности, алгоритмы дискретной оптимизации, верификация ПО, архитектура информационных систем. URL: https://discopal.ispras.ru/stas.

Stanislav Alexandrovich FOMIN — leading programmer. Research interests: complexity theory, discrete optimization algorithms, information systems architecture.

Борис Владимирович КОНДРАТЬЕВ; сфера научных интересов: безопасность информации, защита информации от несанкционированного доступа и утечки по техническим каналам, построение информационных систем в защищённом исполнении, сертификация программного обеспечения по требованиям безопасности информации.

Boris Vladimirovich KONDRAT'EV, research interests: information security, protection of information from unauthorized access and leakage through technical channels, building information systems in a secure design, certification of software for information security requirements.