(19) Евразийское патентное ведомство

(11) 046373 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2024.03.06

(21) Номер заявки
202293340

(22) Дата подачи заявки
2021.06.18

(51) Int. Cl. ***G06F 40/177*** (2020.01)
G06F 40/216 (2020.01)

(54) СПОСОБ И КОМПЬЮТЕРНАЯ СИСТЕМА ДЛЯ УЛУЧШЕННОЙ ОБРАБОТКИ ТАБЛИЦ

(31) **20180937.3**
(32) **2020.06.18**
(33) **EP**
(43) **2023.06.05**
(86) **PCT/US2021/037998**
(87) **WO 2021/257939 2021.12.23**

(71)(73) Заявитель и патентовладелец:
МОРНИНГСТАР ИНК. (US)

(72) Изобретатель:
Котвал Ваибхав, Дешпанде Свапнил, Ядав Картик, Гавад Тушар, Никхил Суле (IN), Шарик Ахмад (US)

(74) Представитель:
Кузнецова С.А. (RU)

(56) NATALIYA LE VINE ET AL: "Identifying Table Structure in Documents using Conditional Generative Adversarial Networks", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 13 January 2020 (2020-01-13), XP081580115, abstract sections 1.3; 1.4 sections 2; 2.1; 2.1.1 section 4

MINGHAO LI ET AL: "TableBank: Table Benchmark for Image-based Table Detection and Recognition", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 5 March 2019 (2019-03-05), XP081129653, abstract section 2.2 sections 3; 3.1; 3.2; figure 3

A SAKILA ET AL: "Image Enhancement using Morphological Operations", INTERNATIONAL JOURNAL OF RECENT RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY, vol. 3, no. 2, 15 April 2017 (2017-04-15), pages 685-698, XP055744061, ISSN: 2395-1990 abstract sections 3.2; 3.5.1; 3.5.2

SIAVASH ARJOMAND BIGDELI ET AL: "Image Restoration using Autoencoding Priors", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 29 March 2017 (2017-03-29), XP080752870, DOI: 10.5220/0006532100330044 abstract

ZHAO GUOPING ET AL: "Skip-Connected Deep Convolutional Autoencoder for Restoration of Document Images", 2018 24TH INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION (ICPR), IEEE, 20 August 2018 (2018-08-20), pages 2935-2940, XP033459845, DOI: 10.1109/ICPR.2018.8546199 abstract

(57) В первом аспекте настоящее изобретение относится к реализуемому на компьютере способу улучшенной обработки таблицы без сетки. Во втором и третьем аспектах настоящее изобретение относится к компьютерной системе и компьютерному программному продукту для улучшенной обработки таблицы без сетки. В четвертом аспекте настоящее изобретение относится к применению любого способа, системы или продукта для разбора документов.

046373 B1

046373 B1

Область техники, к которой относится изобретение

Настоящее изобретение относится к способу, системе и компьютерному программному продукту для улучшенной обработки таблиц, а также их использования для разбора документов.

Уровень техники изобретения

Способы, системы и компьютерные программные продукты для обработки таблиц известны из уровня техники.

Такая обработка таблиц, в частности, полезна для разбора документов с фиксированным макетом, таких как PDF-документы. Разбор таблиц из PDF-документов в виде изображения или без изображения может быть сложной задачей. Различные инструменты разбора PDF с открытым исходным кодом и лицензированные, доступные на рынке, борются за точную обработку таблиц.

Таблицы без сеток, разделяющих столбцы и строки, т.е. таблицы без сетки, особенно сложны для инструментов разбора. Документы, содержащие такие таблицы без сетки, часто упоминают в литературе как полуструктурированные документы. Данные документы сложны для инструментов разбора, поскольку существующие инструменты содержат плохую технологию автоматического обнаружения. В результате аналитику приходится вручную выбирать и извлекать таблицы при разборе документа. Кроме того, существующие инструменты разбора не способны извлекать данные или информацию, содержащиеся в нескольких линиях строк или объединенных ячейках внутри таблицы.

В US 20200089946 описан инструмент для извлечения табличных данных из электронных документов. Данный инструмент генерирует информацию о структуре сетки табличных данных. Эта структура сетки объединена с текстом, связанным с табличными данными, чтобы получить таблицу с сеткой. Авторы настоящего изобретения отмечают, что предоставленный инструмент US '946 является неточным, особенно при генерировании сеток для таблиц, содержащих только горизонтальные или вертикальные линии. Более того, этот инструмент не работает при генерировании сеток для таблиц без сеток.

В данной области техники сохраняется необходимость в улучшенной обработке таблиц и особенно таблиц без сетки, а также в улучшенном разборе документов, содержащих эти таблицы.

Настоящее изобретение направлено на решение по меньшей мере некоторых технических проблем, связанных со способами, системами и компьютерными программными продуктами, известными в данной области техники.

Сущность изобретения

В первом аспекте настоящее изобретение относится к способу обработки таблиц без сетки согласно п.1.

Во втором аспекте настоящее изобретение относится к компьютерной системе обработки таблиц без сетки согласно п.12.

В третьем аспекте настоящее изобретение относится к компьютерному программному продукту для обработки таблиц без сетки согласно п.13.

В четвертом аспекте настоящее изобретение относится к использованию любого способа, системы или компьютерного программного продукта для разбора документов согласно п.14.

Настоящее изобретение является целесообразным, поскольку оно позволяет легко и точно обрабатывать таблицы без сетки. В дополнение к этому полуструктурированные документы, содержащие такие таблицы без сетки, легче разбираются. Дополнительные преимущества настоящего изобретения обсуждены ниже в описании, примерах и фигурах.

Предпочтительные варианты осуществления настоящего изобретения обсуждены в пп.2-12 и 15, а также в описании, примерах и фигурах.

Описание графических материалов

На фиг. 1 показан схематический обзор общего рабочего процесса настоящего изобретения, включающий обработку таблиц без сетки и связанных с ними документов.

На фиг. 2-4 показано прогнозирование локализации таблицы без сетки в документе с использованием модели глубокого обучения, в частности сверточной нейронной сети.

На фиг. 5-10 показано прогнозирование сетки таблицы для извлеченной таблицы без сетки с использованием генеративно-состязательной сети, в частности условной GAN.

На фиг. 11-15 показаны предпочтительные этапы обработки изображения до и после расширения для улучшения обработки таблиц без сетки.

Подробное описание изобретения

Настоящее изобретение относится к реализуемому на компьютере способу, компьютерной системе и компьютерному программному продукту для улучшенной обработки таблиц без сетки, а также к использованию любого из способов, систем или продуктов для разбора документов. Ниже настоящее изобретение будет подробно описано, предпочтительные варианты осуществления обсуждены и настоящее изобретение будет проиллюстрировано с помощью не ограничивающих примеров.

Если не указано иное, то все термины, используемые в описании настоящего изобретения, в том числе технические и научные термины, имеют такое значение, которое обычно понимается специалистом в области техники, к которой относится настоящее изобретение. Определения терминов включены в качестве дополнительного руководства для лучшего понимания идей настоящего изобретения. В рамках

данного документа следующие термины имеют следующие значения.

Используемая в данном документе форма единственного числа относится как к одиночным, так и к множественным упоминаемым объектам, если из контекста явным образом не следует иное. Например, термин "отсек" относится к одному или более чем к одному отсеку.

Термины "содержать", "содержащий", "содержит", "состоит из" в рамках данного документа являются синонимами терминов "включать", "включающий", "включает" и представляют собой инклюзивные, или неограничивающие, термины, которые определяют наличие того, что следует за ними, например компонента, и не исключают или не препятствуют наличию дополнительных неперечисленных признаков, элементов, этапов, известных в данной области техники или описанных в данном документе.

Поскольку термин "один или более" или "по меньшей мере один", как, например, "один или более элементов" или "по меньшей мере один элемент" из группы элементов, ясен сам по себе, для дополнительного примера, этот термин включает, среди прочего, ссылку на любой один из указанных элементов или любые два или более указанных элементов, как, например, на любые ≥3, ≥4, ≥5, ≥6 или ≥7 и т.д. указанных элементов вплоть до всех указанных элементов.

Если не указано иное, то все термины, используемые в описании настоящего изобретения, в том числе технические и научные термины, имеют такое значение, которое обычно понимается специалистом в области техники, к которой относится настоящее изобретение. Определения терминов, используемых в описании, включены как дополнительное руководство для лучшего понимания идей настоящего изобретения. Термины или определения, используемые в данном документе, предоставлены исключительно для помощи в понимании настоящего изобретения.

Отсылка к "одному варианту осуществления" или "варианту осуществления" повсюду в данном описании означает, что конкретный признак, конструкция или характеристика, описанная в связи с этим вариантом осуществления, включена в по меньшей мере один вариант осуществления настоящего изобретения. Таким образом, все фразы "в одном варианте осуществления" и "в варианте осуществления", появляющиеся в разных местах повсюду в данном описании, необязательно относятся, но могут относиться, к одному и тому же варианту осуществления. Кроме того, конкретные признаки, конструкции или характеристики могут комбинироваться любым подходящим образом, как очевидно специалисту в данной области техники их данного описания, в одном или более вариантах осуществления. Кроме того, несмотря на то, что некоторые варианты осуществления, описанные в данном документе, содержат некоторые, но не все, признаки, содержащиеся в других вариантах осуществления, комбинации признаков разных вариантов осуществления подразумеваются находящимися в пределах объема настоящего изобретения и образуют другие варианты осуществления, как понятно специалистам в данной области техники. Например, в нижеследующей формуле изобретения любые из заявленных вариантов осуществления могут использоваться в любой комбинации.

Кроме того, термины "первый", "второй", "третий" и т.д. в описании и формуле изобретения используются для проведения различий между подобными элементами и необязательно - для описания последовательного или хронологического порядка, если это не указано. Следует понимать, что используемые таким образом термины являются взаимозаменяемыми при соответствующих обстоятельствах, и что варианты осуществления настоящего изобретения, описанные в данном документе, выполнены с возможностью эксплуатации в последовательностях, отличающихся от описанных или изображенных в данном документе.

В первом аспекте настоящее изобретение относится к реализуемому на компьютере способу улучшенной обработки таблиц. Способ предпочтительно включает этап предоставления электронного документа с фиксированным макетом, содержащего таблицу без сетки. Способ предпочтительно включает этап обучения модели глубокого обучения (DLM) на обучающих данных, относящихся к множеству электронных документов, связанных с обучением. Предпочтительно каждый из множества электронных документов, связанных с обучением, содержит обучающую таблицу без сетки и связанную с ней метку, указывающую на ограничительную рамку обучающей таблицы. Способ предпочтительно включает этап определения ограничительной рамки для таблицы без сетки в указанном электронном документе с фиксированным макетом посредством обученной DLM. Способ предпочтительно включает этап извлечения изображения таблицы без сетки из указанного электронного документа с фиксированным макетом на основе определенной ограничительной рамки. Способ предпочтительно включает этап обработки извлеченного изображения по меньшей мере посредством выполнения этапа расширения. Способ предпочтительно включает этап обучения условной генеративно-состязательной сети (cGAN), содержащей нейронную сеть-генератор (GNN) и нейронную сеть-дискриминатор (DNN). Предпочтительно, при этом cGAN обучается на обучающих данных, содержащих набор реальных пар. Предпочтительно, при этом каждый из указанного набора реальных пар содержит относящееся к обучению расширенное изображение таблицы без сетки и соответствующее изображение сетки таблицы. Предпочтительно, при этом обучение cGAN включает множество этапов обучения. Предпочтительно, при этом каждый из указанных этапов обучения включает один из этапов a) или b) и этап c):

a) предоставление DNN реальной пары, полученной из набора реальных пар;

b) предоставление DNN сгенерированной пары, содержащей изображение пригодной поддельной

сетки и соответствующее относящееся к обучению расширенное изображение таблицы без сетки, полученное из набора реальных пар, при этом изображение пригодной поддельной сетки получают посредством модификации с помощью GNN соответствующего относящегося к обучению расширенного изображения таблицы без сетки с использованием случайного вектора данных;

c) определение с помощью DNN указания того, является ли предоставленная DNN реальная или сгенерированная пара парой, сгенерированной GNN.

Предпочтительно, при этом обучение cGAN включает множество циклов итеративного обучения GNN и DNN. Предпочтительно, при этом каждый из указанного множества циклов итеративного обучения GNN и DNN содержит по меньшей мере один из указанного множества этапов обучения. Предпочтительно, при этом во время каждого цикла обучения соответствующую функцию потерь, связанную с GNN или DNN, минимизируют до тех пор, пока не будет выполнен связанный предикат цикла стабильности. Предпочтительно, при этом cGAN обучается посредством минимизации комбинированных потерь функций потерь, связанных с GNN и DNN, до тех пор, пока не будет выполнен предикат стабильности комбинированных потерь. Способ предпочтительно дополнительно включает этап определения изображения сетки таблицы для обработанного извлеченного изображения посредством обученной GNN, содержащей обученную cGAN. Способ предпочтительно дополнительно включает этап комбинирования определенного изображения сетки таблицы и таблицы без сетки из предоставленного электронного документа с фиксированным макетом в изображении с сеткой таблицы без сетки.

Во втором аспекте настоящее изобретение относится к компьютерной системе для улучшенной обработки таблиц. Предпочтительно, при этом компьютерная система выполнена с возможностью выполнения реализуемого на компьютере способа согласно первому аспекту настоящего изобретения.

В третьем аспекте настоящее изобретение относится к компьютерному программному продукту для улучшенной обработки таблиц. Предпочтительно, при этом компьютерный программный продукт содержит команды, которые, когда компьютер выполняет компьютерный программный продукт, заставляют компьютер выполнять реализуемый на компьютере способ согласно первому аспекту настоящего изобретения. Предпочтительно, при этом компьютер представляет собой компьютерную систему согласно второму аспекту настоящего изобретения.

В четвертом аспекте настоящее изобретение относится к использованию реализуемого на компьютере способа улучшенной обработки таблиц первого аспекта настоящего изобретения, компьютерная система согласно второму аспекту настоящего изобретения или компьютерный программный продукт согласно третьему аспекту настоящего изобретения для генерирования разобранного документа из электронного документа с фиксированным макетом.

Настоящее изобретение предоставляет реализуемый на компьютере способ, компьютерную систему и компьютерный программный продукт для улучшенной обработки таблиц, а также использование любого из способа, системы или продукта для генерирования разобранного документа из электронного документа с фиксированным макетом. Специалист в данной области техники оценит, что способ реализован в компьютерном программном продукте и выполнен с использованием компьютерной системы. Специалисту в данной области техники также ясно, что улучшенная обработка таблиц может быть использована для разбора таблиц из документов. Поэтому далее четыре аспекта настоящего изобретения рассматриваются вместе.

Предметом настоящего изобретения является улучшение обработки таблиц и, в частности, улучшение обработки таблиц без сетки. Хотя предоставленное решение способно улучшить обработку любой таблицы, содержащей или не имеющей какой-либо структуры сетки, известной в данной области техники, настоящее изобретение особенно направлено на улучшенную обработку таблиц без сетки. Такие таблицы, в частности, трудно обнаружить в электронных документах с фиксированным макетом. Кроме того, такую информацию, как иерархия, содержащуюся в ней, трудно извлечь. Следовательно, предмет настоящего изобретения более конкретно направлен на предоставление способа, системы и компьютерного программного продукта, позволяющих удобно, эффективно и надежно извлекать табличные данные из электронных документов при сохранении их формата и структуры (или иерархии). Улучшенная обработка согласно настоящему изобретению таблиц без сетки может быть преимущественно использована в разных областях, таких как цифровое извлечение символов (DCE), улучшение изображения, оптическое распознавание символов (OCR), анализ макета документа, обнаружение полей, разбор полей и т.д.

Предметом настоящего изобретения является генерирование разобранного документа из электронного документа с фиксированным макетом. В связи с этим, определенное изображение сетки таблицы и таблица без сетки из предоставленного электронного документа с фиксированным макетом скомбинированы в изображении с сеткой таблицы без сетки. Данное изображение с сеткой таблицы без сетки может быть предоставлено в виде отдельного изображения, а может и не быть предоставлено. Данное изображение с сеткой таблицы без сетки может быть наложено на электронный документ с фиксированным макетом или его изображение, а может и не быть наложено. Предпочтительно, генерирование разобранного документа из электронного документа с фиксированным макетом включает этап выполнения OCR текста, связанного с табличными данными, из изображения с сеткой таблицы без сетки.

"Разбор" в рамках данного документа относится к термину, известному из уровня техники, который

предпочтительно следует понимать как анализ и/или разделение потока данных, например, электронного документа с фиксированным макетом, на более легко обрабатываемые компоненты. Это направлено на облегчение выполнения какого-либо преобразования, например, OCR, в потоке данных.

"Табличные данные" в рамках данного документа относятся к термину, известному из уровня техники, который предпочтительно следует понимать как любое представление данных в форме надлежащего и четко определенного формата и структуры (или иерархии), чтобы улучшить визуальное представление, интерпретацию и извлечение данных. Табличные данные помогают поддерживать иерархию данных, а также позволяют классифицировать представленные в них данные. Табличные данные могут включать конфигурацию структурированных данных, такие как структура сетки, таблица и т.д. В случае таблицы без сетки конфигурации структурированных данных невидимы. Необязательно табличные данные могут содержать блок-схему.

"Электронный документ" в рамках данного документа относится к термину, известному из уровня техники, который предпочтительно следует понимать как любой электронный носитель, который может содержать в себе одну или несколько форм содержимого. Такой электронный документ может существовать в разных форматах, таких как файл формата переносимых документов (имеющий расширение .PDF), документ, сгенерированный с использованием программного обеспечения для обработки текстов (например, файл с расширением .doc или .docx), веб-текстовый документ (расширение .HTML или .htm), документ adobe postscript (расширение .ps) и т.д. Кроме того, электронный документ может содержать или не содержать разные формы содержимого, например, конфигурация неструктурированных данных, таких как текстовые данные, изображения и т.д., и конфигурация структурированных данных, таких как табличные данные, гистограмма и т.д. Более того, электронный документ может содержать или не содержать единственные табличные данные (такие как единственная таблица) или множество табличных данных.

"Электронный документ с фиксированным макетом" в рамках данного документа относится к термину, известному из уровня техники, который предпочтительно следует понимать как любой электронный носитель, который может содержать в себе одну или несколько форм содержимого и в котором одна или более форм содержимого имеют фиксированный макет. Такой электронный документ может существовать в разных форматах, таких как файл формата переносимых документов с возможностью поиска или без него (имеющий расширение .PDF), растровый файл (имеющий расширение .bmp), файл формата переносимой сетевой графики (имеющий расширение .png), файл формата объединенной группы экспертов по фотографии (имеющий расширение jpeg), теговый формат файла изображения (имеющий расширение .tiff) и т.д. Предпочтительно электронный документ с фиксированным макетом представляет собой PDF-документ, коротко pdf. Такой PDF широко используется и обеспечивает простоту обмена (например, передачу и получение) табличных данных между разными лицами, а также удобную интерпретацию и извлечение информации из табличных данных. Однако, когда требуется извлечь табличные данные из pdf, например, путем копирования табличных данных из pdf, табличные данные теряют свой формат и структуру (или иерархию), тем самым делая извлеченные данные бесполезными.

Для извлечения изображения таблицы без сетки из предоставленного электронного документа с фиксированным макетом простой вариант осуществления настоящего изобретения предоставляет обучение модели глубокого обучения (DLM) на обучающих данных, относящихся к множеству электронных документов, связанных с обучением. Предпочтительно, при этом каждый из указанного множества электронных документов, связанных с обучением, содержит обучающую таблицу без сетки и связанную с ней метку, указывающую на ограничительную рамку обучающей таблицы.

Преимущественно настоящее изобретение позволяет автоматически определять табличные данные из электронного документа, тем самым, устраняя проблемы, связанные с ручным выбором табличных данных пользователем. Таким образом, время и усилия, необходимые для ручного выбора данных, могут быть сокращены, тем самым повышая эффективность работы пользователя. Кроме того, настоящее изобретение позволяет извлекать таблицы без сетки из электронных документов с фиксированным макетом таким образом, что иерархия табличных данных сохраняется после извлечения. Будет понятно, что сохранение иерархии табличных данных после извлечения позволяет сохранить целостность информации, представленной в табличных данных, тем самым обеспечивая удобное, надежное и эффективное извлечение и интерпретацию табличных данных.

"Глубокое обучение" в рамках данного документа относится к термину, известному из уровня техники, который предпочтительно следует понимать как методы машинного обучения, которые включают сети (такие как искусственные нейронные сети (ANN), рекуррентные нейронные сети (RNN), сверточные нейронные сети (CNN) и т.д.) узлов (таких как искусственные нейроны), способных к полуконтролируемому обучению или контролируемому обучению на основе образцов электронных документов, содержащих различные формы табличных данных в них. В таком случае DLM может быть обучен обнаруживать местоположение табличных данных в электронном документе, например, путем предоставления множества образцов электронных документов, связанных с обучением, содержащих обучающую таблицу без сетки и связанную с ней метку, указывающую на ограничительную рамку обучающей таблицы.

"Ограничительная рамка" в рамках данного документа относится к термину, известному из уровня

техники, который предпочтительно следует понимать как прямоугольная форма, которая создается в области электронного документа, содержащей таблицу без сетки, и в которой созданная прямоугольная форма полностью охватывает таблицу без сетки. Будет понятно, что область одной сгенерированной ограничительной рамки соответствует общей области страницы электронного документа, охватываемой единственными табличными данными. Например, область единственной ограничительной рамки может соответствовать области страницы электронного документа, если указанные табличные данные представлены так, чтобы полностью охватить страницу электронного документа. Однако большинство табличных данных, как правило, охватывают меньшие области внутри страниц электронных документов, например, области в пределах 50% страниц электронного документа. Необязательно область единственной ограничительной рамки может распространяться на несколько страниц, если единственные табличные данные существуют на нескольких страницах электронного документа. В таком случае сгенерированная ограничительная рамка будет соответствовать области, большей, чем общая область каждой страницы электронного документа.

Предпочтительно ограничительная рамка содержит координату, соответствующую вершине ограничительной рамки, и координату, соответствующую ширине и высоте ограничительной рамки. Соответственно ограничительная рамка может содержать координаты: tx; ty; tw; th. Альтернативно ограничительная рамка содержит координату, соответствующую каждой из вершин ограничительной рамки. Соответственно ограничительная рамка может содержать координаты: (tx1, ty1); (tx2, ty2); (tx3, ty3); (tx4, ty4).

Для извлечения изображения таблицы без сетки из предоставленного электронного документа с фиксированным макетом простой вариант осуществления настоящего изобретения, кроме того, обеспечивает извлечение изображения таблицы без сетки из указанного электронного документа с фиксированным макетом на основе определенной ограничительной рамки. Предпочтительно извлеченное изображение соответствует изображению, область которого находится внутри ограничительной рамки электронного документа с фиксированным макетом. Извлеченное изображение может быть или не быть связано с одним или более форматами файлов изображений, в том числе без ограничения .jpg, .png, .bmp, .gif и т.д.

Предпочтительно множество электронных документов, связанных с обучением, каждый из которых содержит обучающую таблицу без сетки и связанную с ней метку, указывающую на ограничительную рамку обучающей таблицы, получают путем:

предоставления множества электронных документов с фиксированным макетом, связанных с обучением DLM, каждый из которых содержит таблицу без сетки, в которой указанное множество электронных документов с фиксированным макетом, связанных с обучением DLM, преобразуют в язык разметки;

преобразования множества электронных документов с фиксированным макетом, связанных с обучением, в соответствующие документы на языке разметки;

идентификации в каждом из соответствующих документов на языке разметки тега таблицы, связанного с таблицей без сетки соответствующего электронного документа с фиксированным макетом, связанного с обучением; и

привязки метки, к каждому из множества электронных документов с фиксированным макетом, связанных с обучением, для ограничительной рамки таблицы без сетки, при этом указанная метка основана по меньшей мере частично на соответствующем идентифицированном теге таблицы.

Согласно предпочтительному варианту осуществления множество электронных документов, связанных с обучением, каждый из которых содержит обучающую таблицу без сетки и связанную с ней метку, указывающую на ограничительную рамку обучающей таблицы, получают путем:

предоставления множества PDF-документов, каждый из которых содержит таблицу без сетки;

преобразования множества PDF-документов в HTML-документы;

идентификации, в каждом из HTML-документов, тега таблицы, связанного с таблицей без сетки соответствующего PDF-документа; и

привязки метки, к каждому из множества PDF-документов, для ограничительной рамки таблицы без сетки, при этом указанная метка основана по меньшей мере частично на соответствующем идентифицированном теге таблицы.

"Язык разметки" в рамках данного документа относится к термину, известному из уровня техники, который предпочтительно следует понимать как формат файла, подходящий для аннотирования документа способом, синтаксически отличимым от соответствующего текста, что означает, когда документ обрабатывают для отображения, язык разметки не отображен и его используют для форматирования соответствующего текста. Использование языка разметки может быть ограничено или не обязательно ограничено форматированием соответствующего текста. Язык разметки может содержать или не содержать дополнительные функциональные возможности. Примеры языков разметки включают следующие форматы файлов troff, nroff, TeX, Scribe, GML, SGML, HTML, XML, XHTML, другие приложения на основе XML и т.д. Использование языка разметки для получения тегов таблицы легко реализовать и предоставляет качественную информацию о положении таблицы (без сетки) в электронном документе. Из этой информации о положении легко выводят ограничительную рамку для таблицы (без сетки). Предпочти-

тельно используют HTML в качестве языка разметки. HTML широко используется.

Предпочтительно DLM для определения ограничительной рамки для таблицы без сетки в электронном документе с фиксированным макетом представляет собой одну или более искусственных нейронных сетей (ANN), рекуррентную нейронную сеть (RNN) или сверточную нейронную сеть (CNN). Согласно предпочтительному варианту осуществления DLM представляет собой CNN. Разные архитектуры DLM поддаются обучению для определения ограничительной рамки для таблицы без сетки в электронном документе с фиксированным макетом. Однако большинство методов не устойчивы к экстремальным изменениям точки зрения и фона. Известно, что CNN чрезвычайно надежны к вариациям фона и точки зрения в задачах обнаружения и классификации объектов, что делает их очень подходящими для определения ограничительных рамок таблицы без сетки в электронных документах с фиксированным макетом с изменяемым макетом, таких как научные публикации и т.д.

Предпочтительно функция потерь, связанная с обучением DLM для определения ограничительной рамки для таблицы без сетки в электронном документе с фиксированным макетом по меньшей мере частично основана на функции потерь пересечения над объединением (IOU). IOU можно рассчитать как область пересечения, разделенную на область объединения двух рамок. IOU должно быть между >0 и <1. Чтобы иметь возможность предсказывать ограничительные рамки, нам нужно, чтобы IOU находилось между предсказанной ограничительной рамкой и эталонной ограничительной рамкой, которая составляет приблизительно 1. Более предпочтительно функция потерь, связанная с обучением DLM для определения ограничительной рамки для таблицы без сетки в электронном документе с фиксированным макетом по меньшей мере частично основана на функции потерь бинарной перекрестной энтропии (BCE).

Извлеченное с помощью CNN изображение обрабатывают согласно простому варианту осуществления настоящего изобретения. Предпочтительно обработка изображения, извлеченного с помощью CNN, включает по меньшей мере выполнение этапа расширения. Выполнение расширения извлеченного изображения связывает объекты, например, текстовую информацию без сетки, присутствующую в таблицах без сетки, с требуемой и/или равномерной плотностью пикселей.

Будет понятно, что при выполнении такого морфологического расширения объектов в таблицах без сетки, чтобы указанные объекты имели равномерную плотность пикселей, уменьшались сложности, связанные с обработкой указанных объектов, имеющих различную плотность пикселей. "Морфологическая операция" в рамках данного документа относится к термину, известному из уровня техники, который предпочтительно следует понимать как метод обработки изображений, в котором манипулируют пикселями изображения на основе форм объектов, представленных на изображении. Кроме того, в морфологической операции используют структурный элемент, например, скользящее окно, которое используют для определения форм объектов, и впоследствии можно манипулировать пикселями, связанными с обнаруженными объектами. "Структурный элемент" в рамках данного документа относится к термину, известному из уровня техники, который предпочтительно следует понимать как шаблон, имеющий предварительно определенную форму, например, прямоугольную, который используют для выполнения морфологической операции.

"Расширение" или "морфологическое расширение" в рамках данного документа относится к термину, известному из уровня техники, который предпочтительно следует понимать как операцию, при которой добавляют пиксели к границе объектов, обнаруженных в изображении.

Согласно предпочтительному варианту осуществления обработка изображения, извлеченного с помощью DLM, дополнительно включает, перед указанным этапом расширения, этапы:

преобразования извлеченного изображения в изображение в оттенках серого; и

применения порогового значения к изображению в оттенках серого с помощью адаптивного метода Гаусса;

при этом указанный этап расширения выполняют для указанного порогового изображения в оттенках серого.

Согласно предпочтительному варианту осуществления обработка извлеченного изображения дополнительно включает, после указанного этапа расширения, этапы:

получения контуров расширенных объектов на расширенном изображении с помощью упомянутого этапа расширения;

расширения контура, содержащего изображение; и

применения порогового значения к расширенному контуру, содержащему изображение, с помощью адаптивного метода Гаусса.

Согласно наиболее предпочтительному варианту осуществления обработка извлеченного изображения включает этапы:

преобразования извлеченного изображения в изображение в оттенках серого;

применения порогового значения к изображению в оттенках серого с помощью адаптивного метода Гаусса;

расширения порогового изображения в оттенках серого;

получения контуров расширенных объектов на расширенном пороговом изображении в оттенках серого;

расширения контура, содержащего изображение; и

применения порогового значения к расширенному контуру, содержащему изображение, с помощью адаптивного метода Гаусса.

"Изображение в оттенках серого" в рамках данного документа относится к термину, известному из уровня техники, который предпочтительно следует понимать как монохроматическое изображение, состоящее из пикселей, имеющих разные оттенки серого цвета. Будет понятно, что разные оттенки серого цвета образуются комбинациями черного и белого цветов в различных пропорциях, например, таких как темный оттенок серого цвета, в котором черный цвет находится в максимальной пропорции, в то время как белый цвет находится в минимальной пропорции, и светлый оттенок серого цвета, в котором черный цвет находится в минимальной пропорции, в то время как белый цвет находится в максимальной пропорции. Такое преобразование извлеченного изображения в изображение в оттенках серого сводит к минимуму сложности, которые могут возникнуть во время обработки цветных изображений, например, посредством использования методов обработки изображений на основе компьютерного зрения, например, морфологических операций, таких как расширение и эрозия. Изображения в оттенках серого также связаны с меньшим количеством информации о пикселях изображения по сравнению с изображениями RGB (красный, зеленый и синий), преобразование принятого изображения в изображение в оттенках серого дополнительно способствует обработке, например, ускорению его обработки.

"Пороговое значение" в рамках данного документа относится к термину, известному из уровня техники, который предпочтительно следует понимать как классифицирование пикселей изображения, например, изображения в оттенках серого, на две группы для получения бинарного изображения, содержащего только два цвета, в котором пиксели классифицированы на основе порогового значения интенсивности. Пиксели, интенсивность которых меньше порогового значения интенсивности (например, пиксели, связанные со светлыми оттенками серого), классифицируют в первую группу, а пиксели, интенсивность которых превышает пороговое значение интенсивности (например, пиксели, связанные с темными оттенками серого), классифицируют во вторую группу. В результате изображение в оттенках серого, содержащее пиксели различных оттенков серого, будет преобразовано в пороговое изображение в оттенках серого, содержащее пиксели только черного и белого цветов соответственно, где пиксели, имеющие черный цвет, соответствуют пикселям изображения в оттенках серого для всех значений интенсивности, меньших, чем пороговое значение интенсивности, а пиксели, имеющие белый цвет, соответствуют пикселям изображения в оттенках серого для всех значений интенсивности, превышающих пороговое значение интенсивности. Кроме того, выполнение порогового значения изображения, например, изображения в оттенках серого, сегментирует пороговое изображение в оттенках серого, так что фон порогового изображения в оттенках серого, связанного с пустым пространством, отделен от переднего плана порогового изображения в оттенках серого, связанного с текстом и структурой сетки табличных данных.

"Адаптивный метод Гаусса" в рамках данного документа относится к термину, известному из уровня техники, который предпочтительно следует понимать как метод порогового значения, который использует значения интенсивности пикселей, ближайших к заданному пикселю (например, соседних пикселей), для классификации заданного пикселя в первую группу или вторую группу. Кроме того, адаптивный метод Гаусса учитывает взвешенную сумму значений интенсивности пикселей, ближайших к заданному пикселю, и расстояние ближайших пикселей от заданного пикселя для классификации заданного пикселя в первую группу или вторую группу. Более того, адаптивный метод Гаусса сегментирует изображение, например, изображение в оттенках серого, на множество подизображений на основе изменений в изображении, например, изображения в оттенках серого (например, на основе изменений в фоне изображения в оттенках серого) и впоследствии рассматривает динамические пороговые значения интенсивности для каждого подизображения на основе средневзвешенных значений интенсивности пикселей, ближайших к заданному пикселю в подизображений, и постоянного значения. Будет понятно, что выполнение порогового значения изображения, например, изображения в оттенках серого, путем учета динамических пороговых значений интенсивности, т.е. с использованием адаптивного метода Гаусса, позволяет улучшить четкость и точность, связанные с пороговым значением изображения.

Согласно предпочтительному варианту осуществления обработка изображения, извлеченного с помощью DLM, дополнительно включает, после по меньшей мере указанного этапа расширения, этап эрозии. Наиболее предпочтительно обработка изображения, извлеченного с помощью DLM, дополнительно включает, после каждого выполненного этапа расширения, этап эрозии.

"Эрозия" или "морфологическая эрозия" в рамках данного документа относится к термину, известному из уровня техники, который предпочтительно следует понимать как операция, при которой пиксели удаляют с границы объектов, обнаруженных в изображении, например, путем удаления самого внешнего слоя пикселей, связанного с объектами в изображении. За счет последующего морфологического расширения и размывания объектов в изображении таблицы без сетки дополнительно уменьшают сложности, связанные с обработкой указанных объектов.

Для определения изображения сетки таблицы для изображения, извлеченного и обработанного DLM, простой вариант осуществления настоящего изобретения дополнительно обеспечивает обучение условной генеративно-состязательной сети (cGAN), содержащей нейронную сеть-генератор (GNN) и

нейронную сеть-дискриминатор (DNN).

Генеративно-состязательные сети (GANs) представляют собой ветвь неконтролируемого машинного обучения и реализованы системой из двух нейронных сетей, т.е. GNN и DNN, конкурирующих друг с другом в рамках игры с нулевой суммой. Создаются две нейронные сети и выполняется обучение (т.е. сети обучаются) посредством их взаимной конкуренции. Первая нейронная сеть реализована как система-генератор и называется нейронной сетью-генератором (GNN). Вторая нейронная сеть реализована как система-дискриминатор и называется нейронной сетью-дискриминатором (DNN). GNN начинает со случайного ввода и пытается сгенерировать синтетические или поддельные изображения. DNN получает реальные аутентичные изображения вместе с синтетическими изображениями от GNN. Соответственно DNN выводит бинарное решение, указывающее прогноз относительно того, является ли изображение, полученное от GNN, настоящим или синтетическим (т.е. поддельным). Поскольку DNN улучшает различение между настоящим изображением и синтетическими изображениями, то GNN улучшает генерирование изображений для обмана DNN. GNN и DNN итеративно обучаются путем минимизации соответствующей связанной функции потерь. Равновесие, например, достигается, когда GNN больше не может обманывать DNN. Никакие знания в предметной области не должны быть явно включены. Особенности изображения изучаются автоматически. Функция потерь для GNN изучена и предварительно не определена. В условных GANs (CGANs) вместо генерирования выборки из случайного ввода GNN генерирует выходное изображение, обусловленное входным изображением. Затем такую систему можно обучить для изучения отображений из пространства входного изображения в пространство выходного или сгенерированного изображения. Процесс обучения заключается в изучении оптимального набора значений множества параметров, определяющих отображение. Другими словами, отображение можно рассматривать как функциональное отображение с регулируемыми параметрами, которые изучают на основе набора обучающих выборок.

Предпочтительно cGAN представляет собой GAN попиксельную GAN (pix2pix). Модель pix2pix представляет собой тип cGAN, где генерирование выходного изображения зависит от входных данных, в данном случае это обработанное, т.е. расширенное изображение изображения, извлеченного с помощью DLM. DNN предоставляют как с таким обработанным изображением, так и с целевым изображением, и она должна определить, является ли цель правдоподобным преобразованием обработанного изображения. GNN обучается через состязательные потери, что побуждает GNN генерировать правдоподобные изображения в целевой предметной области. GNN также обновляют с помощью потерь L1, измеренных между сгенерированным изображением и ожидаемым выходным изображением. Эта дополнительная потеря побуждает модель GNN создавать правдоподобные трансляции исходного изображения. Более предпочтительно, при этом GNN pix2pix GAN представляет собой CNN. Еще более предпочтительно, при этом GNN pix2pix GAN представляет собой CNN, содержащую архитектуру U-net. U-net представляет собой архитектуру CNN для быстрой и точной сегментации изображений.

Предпочтительно cGAN обучена на обучающих данных, содержащих набор реальных пар. Предпочтительно, при этом каждый набор из указанного набора реальных пар содержит относящееся к обучению расширенное изображение таблицы без сетки и соответствующее изображение сетки таблицы. Предпочтительно, при этом обучение cGAN включает множество этапов обучения. Предпочтительно, при этом каждый из указанных этапов обучения включает один из этапов a) или b) и этап c):

a) предоставление DNN реальной пары, полученной из набора реальных пар;

b) предоставление DNN сгенерированной пары, содержащей изображение пригодной поддельной сетки и соответствующее относящееся к обучению расширенное изображение таблицы без сетки, полученное из набора реальных пар, при этом изображение пригодной поддельной сетки получают посредством модификации с помощью GNN соответствующего относящегося к обучению расширенного изображения таблицы без сетки с использованием случайного вектора данных;

c) определение с помощью DNN указания того, является ли предоставленная DNN реальная или сгенерированная пара парой, сгенерированной GNN.

Предпочтительно, при этом обучение cGAN включает множество циклов итеративного обучения GNN и DNN. Предпочтительно, при этом каждый из указанного множества циклов итеративного обучения GNN и DNN содержит по меньшей мере один из указанного множества этапов обучения. Предпочтительно, при этом во время каждого цикла обучения соответствующую функцию потерь, связанную с GNN или DNN, минимизируют до тех пор, пока не будет выполнен связанный предикат цикла стабильности. Предпочтительно, при этом cGAN обучается посредством минимизации комбинированных потерь функций потерь, связанных с GNN и DNN, до тех пор, пока не будет выполнен предикат стабильности комбинированных потерь.

Предпочтительно каждое из относящихся к обучению расширенных изображений таблицы без сетки указанного набора реальных пар получают путем расширения изображения таблицы без сетки.

Аналогично изображению, извлеченному с помощью DLM, как обсуждалось выше, производительность настоящего изобретения улучшается, в частности производительность cGAN, когда относящиеся к обучению расширенные изображения таблицы без сетки получаются, как описано ниже.

Более предпочтительно каждое из относящихся к обучению расширенных изображений таблицы

без сетки указанного набора реальных пар получают путем:

преобразования изображения таблицы без сетки в изображение в оттенках серого;

применения порогового значения к изображению в оттенках серого с помощью адаптивного метода Гаусса; и

расширения порогового изображения в оттенках серого.

Еще более предпочтительно каждое из относящихся к обучению расширенных изображений таблицы без сетки указанного набора реальных пар получают путем:

преобразования изображения таблицы без сетки в изображение в оттенках серого;

применения порогового значения к изображению в оттенках серого с помощью адаптивного метода Гаусса;

расширения порогового изображения в оттенках серого;

получения контуров расширенных объектов на расширенном пороговом изображении в оттенках серого;

расширения контура, содержащего изображение; и

применения порогового значения к расширенному контуру, содержащему изображение, с помощью адаптивного метода Гаусса.

Предпочтительно набор из реальных пар, каждая из которых содержит относящееся к обучению расширенное изображение таблицы без сетки и соответствующее изображение сетки таблицы, получают путем:

предоставления множества электронных документов с фиксированным макетом, связанных с обучением cGAN, каждый из которых содержит таблицу без сетки, в которой указанное множество электронных документов с фиксированным макетом, связанных с обучением cGAN, преобразуют в язык разметки;

преобразования множества электронных документов с фиксированным макетом, связанных с обучением, в соответствующие документы на языке разметки;

идентификации в каждом из соответствующих документов на языке разметки тега таблицы, связанного с таблицей без сетки соответствующего электронного документа с фиксированным макетом, связанного с обучением; и

определения для каждой из таблиц без сетки изображения сетки таблицы, основанного по меньшей мере частично на соответствующем идентифицированном теге таблицы;

получения из каждого из указанного множества электронных документов с фиксированным макетом, связанных с обучением cGAN, изображения таблицы без сетки; и

расширения каждого из полученных изображений таблиц без сетки.

Более предпочтительно изображение таблицы без сетки получают из каждого из указанных предоставленных pdf-документов путем:

определения ограничительной рамки для таблицы без сетки в каждом из электронных документов с фиксированным макетом, связанных с обучением cGAN, с помощью обученной DLM; и

извлечения из каждого из указанных электронных документов с фиксированным макетом, связанных с обучением cGAN, изображения таблицы без сетки на основе определенной соответствующей ограничительной рамки.

Согласно предпочтительному варианту осуществления набор из реальных пар, каждая из которых содержит относящееся к обучению расширенное изображение таблицы без сетки и соответствующее изображение сетки таблицы, получают путем:

предоставления множества pdf-документов, каждый из которых содержит таблицу без сетки;

преобразования множества pdf-документов в html-документы;

идентификации, в каждом из html-документов, тега таблицы, связанного с таблицей без сетки соответствующего pdf-документа;

определения для каждой из таблиц без сетки изображения сетки таблицы, основанного по меньшей мере частично на соответствующем идентифицированном теге таблицы;

получения из каждого из указанных pdf-документов изображения таблицы без сетки; и

расширения каждого из полученных изображений таблиц без сетки.

Согласно дополнительному предпочтительному варианту осуществления изображение таблицы без сетки получают из каждого из указанных предоставленных pdf-документов путем:

определения ограничительной рамки для таблицы без сетки в каждом pdf-документе с помощью обученной DLM; и

извлечения из каждого из указанных pdf-документов изображения таблицы без сетки на основе определенной соответствующей ограничительной рамки.

Аналогично обучающим данным DLM, использование языка разметки для получения тегов таблицы легко реализовано и предоставляет качественную информацию о положениях невидимой сетки таблицы без сетки в электронном документе. Из этой информации о положении невидимой сетки легко выводят изображение сетки таблицы для таблицы без сетки. Предпочтительно используют HTML, который широко используется, в качестве языка разметки.

Для определения изображения с сеткой таблицы без сетки простой вариант осуществления настоящего изобретения дополнительно предоставляет: определение изображения сетки таблицы для обработанного извлеченного изображения с помощью обученной GNN, состоящей из обученной cGAN; и комбинирование определенного изображения сетки таблицы и таблицы без сетки из предоставленного электронного документа с фиксированным макетом в указанном изображении с сеткой. Изображение сетки таблицы и таблица без сетки могут быть объединены для получения указанного изображения с сеткой с использованием любого метода, известного из уровня техники, такого как слияние изображений или наложение одного из изображений на другое. Используемый метод может зависеть или не зависеть от форматов файлов изображений, используемых при внедрении настоящего изобретения.

Как и в случае с любым методом машинного обучения (MLT), качество генерируемых выходных данных в значительной степени зависит от качества обучающих данных, на основе которых обучается модель MLT. В случае обученной cGAN отсутствие обучающих данных может привести или не привести к неадекватным или поврежденным определенным изображениям сетки таблицы. Объединение такого ненадлежащего или поврежденного определенного изображения сетки таблицы и таблицы без сетки в изображение с сеткой может препятствовать или не препятствовать обработке и/или разбору. Для улучшения качества ненадлежащих или поврежденных определенных изображений сетки таблиц могут быть использованы известные в данной области техники методы восстановления изображений.

Согласно предпочтительному варианту осуществления настоящее изобретение дополнительно включает этап обучения искусственной нейронной сети (ANN) на обучающих данных, содержащих обучающие пары, каждая из которых содержит изображение сетки таблицы, относящееся к обучению, и соответствующее поврежденное изображение сетки таблицы. Предпочтительно, при этом ANN представляет собой автокодер. Согласно предпочтительному варианту осуществления настоящее изобретение дополнительно содержит этап определения восстановленного изображения сетки таблицы из поврежденного изображения сетки таблицы посредством обученной ANN, содержащей изображение сетки таблицы, определенное обученной GNN. Предпочтительно, при этом определенное восстановленное изображение сетки таблицы и таблица без сетки из предоставленного электронного документа с фиксированным макетом скомбинированы в изображении с сеткой таблицы без сетки.

Примеры

Настоящее изобретение дополнительно описано посредством следующих неограничивающих примеров, которые дополнительно иллюстрируют настоящее изобретение и не предназначены для ограничения объема настоящего изобретения и их не следует интерпретировать как ограничивающие объем настоящего изобретения.

Пример 1. Разбор документа.

Настоящий пример относится к общему обзору рабочего процесса согласно настоящему изобретению. Ссылка сделана на фиг. 1.

На фиг. 1 показан схематический обзор общего рабочего процесса настоящего изобретения. Рабочий процесс включает обработку таблиц без сетки и связанных с ними документов. Согласно настоящему примеру предоставленный электронный документ (1) с фиксированным макетом, содержащий таблицу без сетки, представляет собой PDF-документ (1'), который преобразован в файл (1") изображения. Данный файл изображения передают обученной CNN (2). Обучение указанной DLM обсуждено в примере 2. С использованием этой обученной CNN извлекают изображение таблицы (3) без сетки, которое затем обрабатывают по меньшей мере путем выполнения этапа (4) расширения.

Обработка извлеченного изображения обсуждена в примере 4. Обработанное извлеченное изображение (5) передают обученной GNN, состоящей из обученной GAN (6). Обучение указанной GAN (6) обсуждено в примере 3. С использованием обученной CNN определяют изображение (7) сетки таблицы для обработанного извлеченного изображения. После этого изображение сетки таблицы накладывают (8) на извлеченное изображение таблицы без сетки для получения изображения с сеткой таблицы (9) без сетки. Данное изображение с сеткой передают инструменту (10) OCR для получения табличной информации в таблице без сетки в текстовом формате (11).

Пример 2. Обучение DLM.

Настоящий пример относится к обучению DLM, в частности CNN. Ссылка сделана на фиг. 2-4.

На фиг. 2 и 3 изображено получение обучающих данных для обучения DLM. Обучающие данные относятся к множеству электронных документов (1), связанных с обучением, каждый из которых содержит обучающую таблицу (12) без сетки и связанную с ней метку (13), указывающую на ограничительную рамку обучающей таблицы. Путем обучения DLM на таких обучающих данных, с помощью обученной DLM можно определить ограничительную рамку (15) для таблицы (12) без сетки в указанном электронном документе (1) с фиксированным макетом.

Согласно настоящему примеру DLM представляет собой CNN, в частности основанную на общедоступном инструменте YOLO. YOLO является мощной нейронной сетью, которая рисует ограничительные рамки вокруг обнаруженных объектов в изображении. Darknet является платформой с открытым исходным кодом, который используется в обучении нейронных сетей и служит основой для YOLO. В настоящем примере сеть YOLOV3 используют для идентификации таблиц без сетки. Для обучения сети

YOLOV3 PDF-документы на основе изображений сначала были преобразованы в выходные данные HTML. Затем содержимое HTML было фрагментировано и разобрано для таблиц. Теги таблицы, например <td>, из исходного кода HTML были обнаружены и окрашены. Это, например, видно на фиг. 2, и позволило координатам таблиц присутствовать в PDF-документах, тем самым образуя обучающий набор, который был передан в YOLOV3. Из этой информации были определены координаты x, y, а также высота и ширина таблицы.

YOLOV3 использует IOU и BCE в качестве функции потерь и использует логистическую регрессию для локализации объекта. IOU можно рассчитать как область пересечения, разделенную на область объединения двух рамок. IOU должно быть >0 и <1. Чтобы иметь возможность предсказывать ограничительные рамки, IOU нужно находиться между предсказанной ограничительной рамкой и эталонной ограничительной рамкой, которая составляет приблизительно 1. Сеть YOLOV3 предсказывает четыре координаты для каждой ограничительной рамки: tx; ty; tw; th. В настоящее время ячейка смещена от верхнего левого угла изображения на (cx, cy), а предшествующая ограничительная рамка имеет ширину и высоту Pw, Ph.

Предсказания сети YOLOV3 соответствуют:

$$bx = \delta(t_x) + c_x$$

$$by = \delta(t_y) + c_y$$

$$b_w = p_w e^{tw}$$

$$b_h = p_h e^{th}$$

Функция потерь сети YOLOV3 показана ниже.

$$\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2]$$
$$+\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i}\right)^2 \right]$$
$$+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left(C_i - \hat{C}_i\right)^2 + \lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{ij}^{noobj} \left(C_i - \hat{C}_i\right)^2$$
$$+ \sum_{i=0}^{S^2} \mathbb{1}_i^{obj} \sum_{c \in classes} \left(p_i(c) - \hat{p}_i(c)\right)^2$$

На фиг. 4 изображены итерации прогнозирования локализации для определения ограничительной рамки таблицы без сетки, например, в изображении. Во время обучения прогнозы (15', 15", 15"') ограничительной рамки итеративно корректируют до тех пор, пока разница с эталонными данными (16) ограничительной рамки таблицы без сетки не будет соответствовать предикату.

Пример 3. Обучение GAN.

Настоящий пример относится к обучению GAN, в частности cGAN. Ссылка сделана на фиг. 5-10.

На фиг. 5 изображен общий обзор архитектуры GAN. GAN реализована системой из двух нейронных сетей, т.е. GNN (14) и DNN (21), конкурирующих друг с другом в рамках игры с нулевой суммой. Создаются две нейронные сети и выполняется обучение (т.е. сети обучаются) посредством их взаимной конкуренции. Первая нейронная сеть реализована как система-генератор и называется нейронной сетью-генератором (GNN). Вторая нейронная сеть реализована как система-дискриминатор и называется нейронной сетью-дискриминатором (DNN). GNN начинает со случайного ввода (17) и пытается сгенерировать синтетические или поддельные изображения (18). DNN получает реальные аутентичные изображения (19; 20) вместе с синтетическими изображениями от GNN. Соответственно DNN выводит бинарное решение (22), указывающее прогноз относительно того, является ли изображение, полученное от GNN, настоящим или поддельным. Поскольку DNN улучшает различение между настоящим изображением и синтетическими изображениями, то GNN улучшает генерирование изображений для обмана DNN. GNN обучается путем минимизации соответствующей связанной функции потерь. Функция потерь для GNN изучена и предварительно не определена. GNN обучается путем минимизации соответствующей связанной функции (23) потерь. GNN и DNN итеративно обучаются путем минимизации соответствующей связанной функции потерь. Равновесие, например, достигается, когда GNN больше не может обманывать DNN. Никакие знания в предметной области не должны быть явно включены. Особенности изображения изучаются автоматически.

Согласно настоящему примеру GAN использует архитектуру PatchGAN. Способ бинарной перекрестной энтропии был использован в качестве функции потерь для DNN. Кроме того, для обучения модели YOLOV3 были смоделированы выходные изображения (пример 2), тем самым предоставляя информа-

цию о координатах в таблице без сетки. Эти изображения предварительно обрабатывают для расширения с использованием компьютерного зрения и передают в DNN как реальные изображения, в то время как выходные изображения GNN передают как поддельные изображения. Обучаемые DNN параметры установлены как ложные, а используемая функция потерь является бинарной перекрестной энтропией. Наконец, все изображения передают в GAN с потерями, установленными на бинарную перекрестную энтропию и среднюю абсолютную ошибку. GAN согласно настоящему примеру представляет собой cGAN и, следовательно, зависит от ее входных данных. Функция потерь всей GAN представляет собой функцию потерь генератора и дискриминатора. Дискриминатор уменьшает реальное и сгенерированное изображение до фрагментов и вычисляет потерю энтропии для каждого фрагмента. Поскольку в настоящем примере используют архитектуру GAN Patch 5, то вся функция потери GAN соответствует:

$$\min_G \max_D V(G,D) = E_{x\sim Px}[\log D(x,y)] + E_{x\sim Px}[\log(\log D(x,G(x,z)))]$$

На фиг. 6-10 изображено прогнозирование сетки таблицы для извлеченной таблицы без сетки с использованием GAN. На фиг. 6 показано расширенное изображение таблицы (3) без сетки. На фиг. 7 показаны выходные данные GAN. Эти выходные данные представляют собой поврежденное изображение (7') сетки. Данное поврежденное изображение сетки может быть восстановлено с помощью метода восстановления изображения. Таким образом, получают восстановленное изображение (7") сетки таблицы. Указанное восстановленное изображение сетки таблицы может быть наложено на восстановленное изображение сетки таблицы, чтобы получить изображение (9) с сеткой таблицы без сетки. Это изображение с сеткой можно разобрать с помощью OCR, чтобы получить читаемую информацию (11) таблицы без сетки.

Пример 4. Расширение изображения.

Настоящий пример относится к обработке до и после расширения изображения. Ссылка сделана на фиг. 11-15.

На фиг. 11-15 изображен предпочтительный этап обработки изображения до и после расширения для улучшения обработки таблиц без сетки. Исходное цветное изображение (24) преобразуют в изображение в оттенках серого. После преобразования к изображению в оттенках серого применяют адаптивное пороговое значение, поскольку статические пороговые значения не используют, и каждый документ сохраняет разный стиль шрифта, размер шрифта, отступы, каллиграфию и т.д. Таким образом, получают пороговое изображение (25) в оттенках серого. На этом изображении выделен шрифт заголовка таблицы. Кроме того, эти этапы помогают динамически определять порог пикселей изображения. Пороговое значение вычисляют с использованием взвешенной суммы соседних значений с использованием окна Гаусса. Затем пороговое изображение в оттенках серого расширяют и размывают. Таким образом, получают расширенное изображение порогового изображения (26) в оттенках серого. Для эрозии и расширения нам нужны два ввода: изображение и структурирующий элемент. Структурирующий элемент называется ядром. При расширении ядро свертывает и выполняет подвыборку по максимальному значению, тем самым расширяя область, охватываемую текстовыми данными. При эрозии ядро выполняет прямо противоположную операцию, то есть свертывает и вычисляет подвыборку по минимальному значению, которая уменьшает внешний вид текстовых данных. После этапа расширения получают контуры (27) расширенных объектов в расширенном изображении на предыдущем этапе расширения. После этого к контуру, содержащему изображение, применяют расширение и к расширенному контуру, содержащему изображение, снова применяют пороговое значение для получения порогового расширенного контура, содержащего изображение (28).

Предполагается, что настоящее изобретение не ограничивается никакой из ранее описанных форм реализации, и что некоторые модификации могут быть добавлены в представленный пример изготовления без пересмотра приложенной формулы изобретения. Способы согласно настоящему изобретению могут быть реализованы многими разными способами без отступления от объема настоящего изобретения.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Реализуемый на компьютере способ улучшенной обработки таблиц, включающий этапы:

предоставления электронного документа с фиксированным макетом, содержащего таблицу без сетки;

обучения модели глубокого обучения (DLM) на обучающих данных, относящихся к множеству электронных документов, связанных с обучением, каждый из которых содержит обучающую таблицу без сетки и связанную с ней метку, указывающую на ограничительную рамку обучающей таблицы;

определения ограничительной рамки для таблицы без сетки в указанном электронном документе с фиксированным макетом с помощью обученной DLM;

извлечения изображения таблицы без сетки из указанного электронного документа с фиксированным макетом на основе определенной ограничительной рамки; и

обработки извлеченного изображения по меньшей мере посредством выполнения этапа расширения;

при этом способ дополнительно включает этапы:

обучения условной генеративно-состязательной сети (cGAN), содержащей нейронную сеть-генератор (GNN) и нейронную сеть-дискриминатор (DNN), на обучающих данных, содержащих набор реальных пар, каждая из которых содержит относящееся к обучению расширенное изображение таблицы без сетки и соответствующее изображение сетки таблицы, причем обучение cGAN включает множество этапов обучения, каждый из которых включает один из этапов a) или b) и этап c):

a) предоставление DNN реальной пары, полученной из набора реальных пар;

b) предоставление DNN сгенерированной пары, содержащей изображение пригодной поддельной сетки и соответствующее относящееся к обучению расширенное изображение таблицы без сетки, полученное из набора реальных пар, при этом изображение пригодной поддельной сетки получают посредством модификации с помощью GNN соответствующего относящегося к обучению расширенного изображения таблицы без сетки с использованием случайного вектора данных;

c) определение с помощью DNN указания того, является ли предоставленная DNN реальная или сгенерированная пара парой, сгенерированной GNN;

при этом обучение cGAN включает множество циклов итеративного обучения GNN и DNN, причем каждый цикл содержит по меньшей мере один из указанного множества этапов обучения, где во время каждого цикла обучения соответствующую функцию потерь, связанную с GNN или DNN, минимизируют до тех пор, пока не будет выполнен связанный предикат цикла стабильности, и при этом cGAN обучается посредством минимизации комбинированных потерь функций потерь, связанных с GNN и DNN, до тех пор, пока не будет выполнен предикат стабильности комбинированных потерь;

определения изображения сетки таблицы для обработанного извлеченного изображения посредством обученной GNN, содержащей обученную cGAN;

комбинирования определенного изображения сетки таблицы и таблицы без сетки из предоставленного электронного документа с фиксированным макетом в изображении с сеткой таблицы без сетки;

при этом множество электронных документов, связанных с обучением, каждый из которых содержит обучающую таблицу без сетки и связанную с ней метку, указывающую на ограничительную рамку обучающей таблицы, получают путем:

предоставления множества PDF-документов, каждый из которых содержит таблицу без сетки;

преобразования множества PDF-документов в HTML-документы;

идентификации, в каждом из HTML-документов, тега таблицы, связанного с таблицей без сетки соответствующего PDF-документа; и

привязки метки, к каждому из множества PDF-документов, для ограничительной рамки таблицы без сетки, при этом указанная метка основана по меньшей мере частично на соответствующем идентифицированном теге таблицы.

2. Способ по п.1, отличающийся тем, что обработка извлеченного изображения дополнительно включает, перед указанным этапом расширения, этапы:

преобразования извлеченного изображения в изображение в оттенках серого; и

применения порогового значения к изображению в оттенках серого с помощью адаптивного метода Гаусса;

при этом указанный этап расширения выполняют для указанного порогового изображения в оттенках серого.

3. Способ по любому из п.1 или 2, отличающийся тем, что обработка извлеченного изображения дополнительно включает, после указанного этапа расширения, этапы:

получения контуров расширенных объектов на расширенном изображении с помощью упомянутого этапа расширения;

расширения контура, содержащего изображение; и

применения порогового значения к расширенному контуру, содержащему изображение, с помощью адаптивного метода Гаусса.

4. Способ по любому из п.2 и 3.

5. Способ по любому из пп.1-4, отличающийся тем, что обработка извлеченного изображения включает, после указанного этапа расширения, этап эрозии.

6. Способ по любому из пп.1-5, отличающийся тем, что способ дополнительно включает этапы:

обучения искусственной нейронной сети (ANN) на обучающих данных, содержащих обучающие пары, каждая из которых содержит изображение сетки таблицы, относящееся к обучению, и соответствующее поврежденное изображение сетки таблицы, предпочтительно, где ANN представляет собой автокодер; и

определения восстановленного изображения сетки таблицы из поврежденного изображения сетки таблицы посредством обученной ANN, содержащей изображение сетки таблицы, определенное обученной GNN;

при этом определенное восстановленное изображение сетки таблицы и таблица без сетки из предоставленного электронного документа с фиксированным макетом скомбинированы в изображении с сеткой таблицы без сетки.

7. Способ по любому из пп.1-6, отличающийся тем, что набор из реальных пар, каждая из которых

содержит относящееся к обучению расширенное изображение таблицы без сетки и соответствующее изображение сетки таблицы, получают путем:

предоставления множества PDF-документов, каждый из которых содержит таблицу без сетки;

преобразования множества PDF-документов в HTML-документы;

идентификации, в каждом из HTML-документов, тега таблицы, связанного с таблицей без сетки соответствующего PDF-документа;

определения для каждой из таблиц без сетки изображения сетки таблицы, основанного по меньшей мере частично на соответствующем идентифицированном теге таблицы;

получения из каждого из указанных PDF-документов изображения таблицы без сетки; и

расширения каждого из полученных изображений таблиц без сетки.

8. Способ по п.7, отличающийся тем, что изображение таблицы без сетки получают из каждого из указанных предоставленных PDF-документов путем:

определения ограничительной рамки для таблицы без сетки в каждом PDF-документе с помощью обученной DLM; и

извлечения из каждого из указанных PDF-документов изображения таблицы без сетки на основе определенной соответствующей ограничительной рамки.

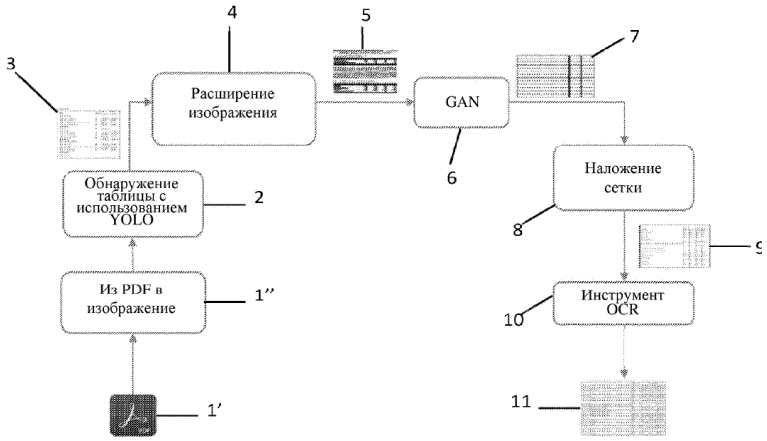
9. Способ по любому из пп.1-8, отличающийся тем, что DLM представляет собой одну или более искусственных нейронных сетей (ANN), рекуррентную нейронную сеть (RNN) или сверточную нейронную сеть (CNN).

10. Способ по любому из пп.1-9, отличающийся тем, что DLM представляет собой CNN.

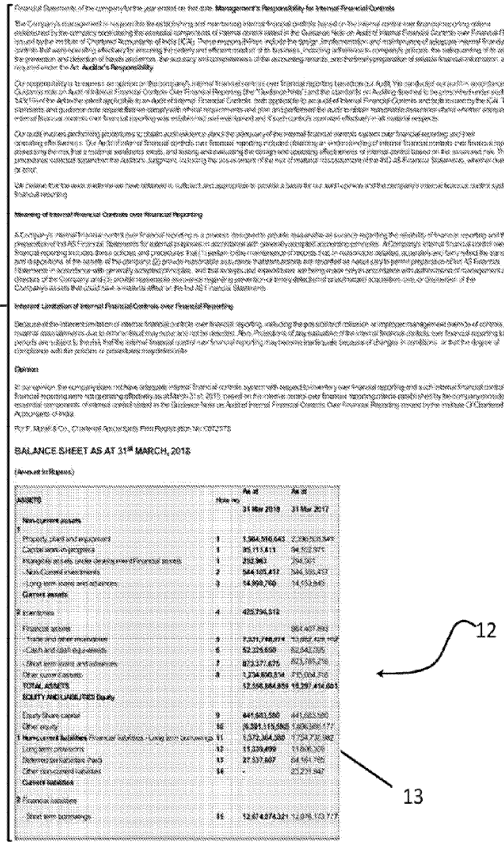
11. Компьютерная система для улучшенной обработки таблиц, при этом компьютерная система выполнена с возможностью выполнения реализуемого на компьютере способа по любому из пп.1-10.

12. Применение реализуемого на компьютере способа по любому из пп.1-10 или компьютерной системы по п.11 для генерирования разобранного документа из электронного документа с фиксированным макетом.

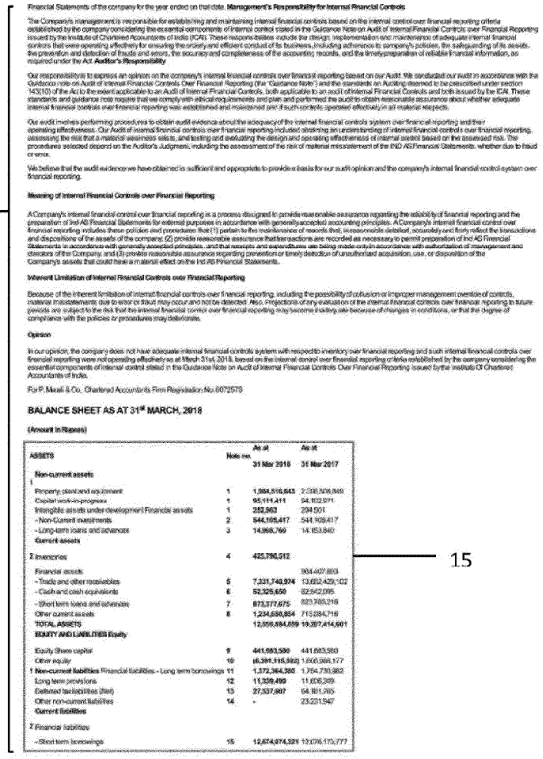
13. Применение по п.12, в котором генерирование разобранного документа из электронного документа с фиксированным макетом включает этап выполнения оптического распознавания символов (OCR) текста, связанного с табличными данными, из изображения с сеткой таблицы без сетки.



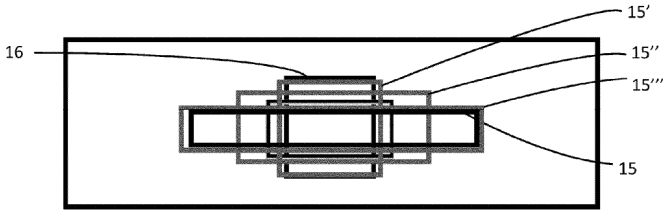
Фиг. 1



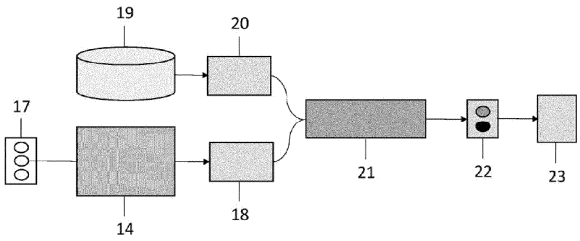
Фиг. 2



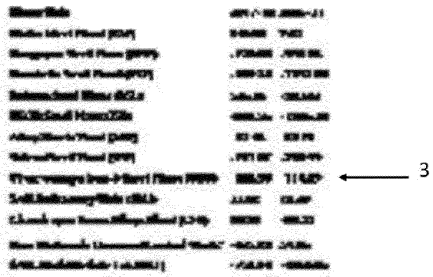
Фиг. 3



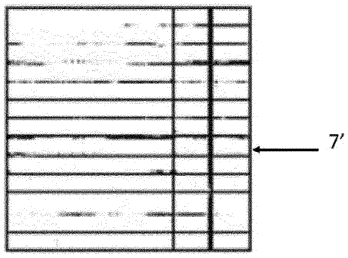
Фиг. 4



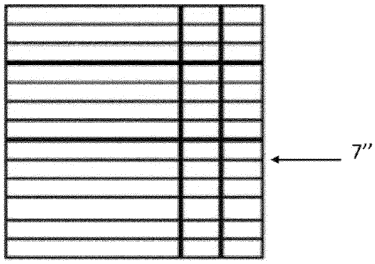
Фиг. 5



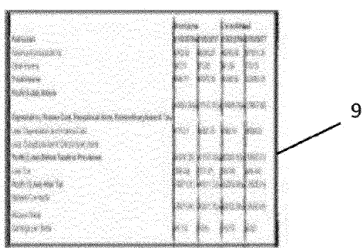
Фиг. 6



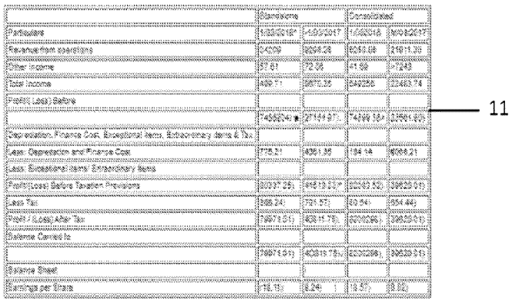
Фиг. 7



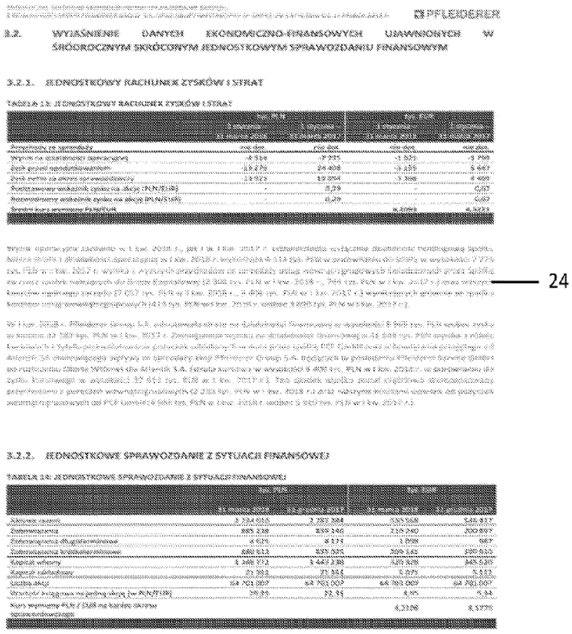
Фиг. 8



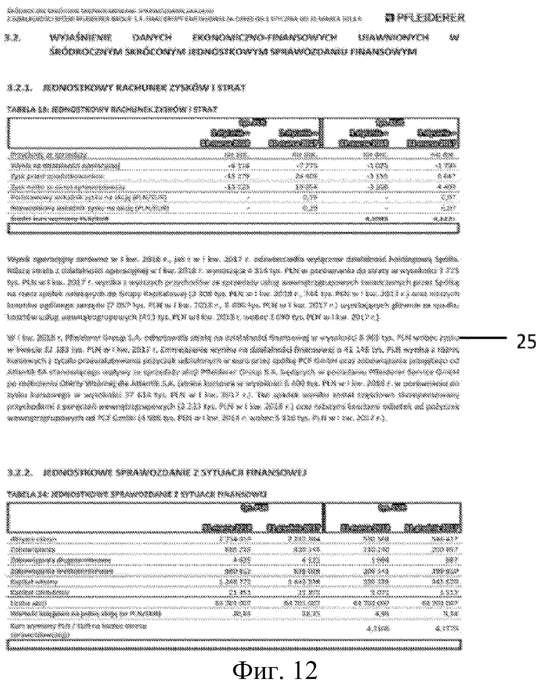
Фиг. 9



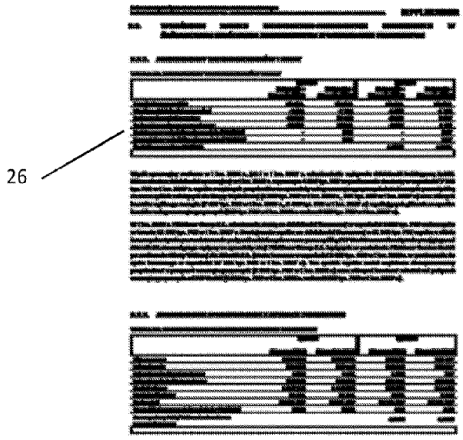
Фиг. 10



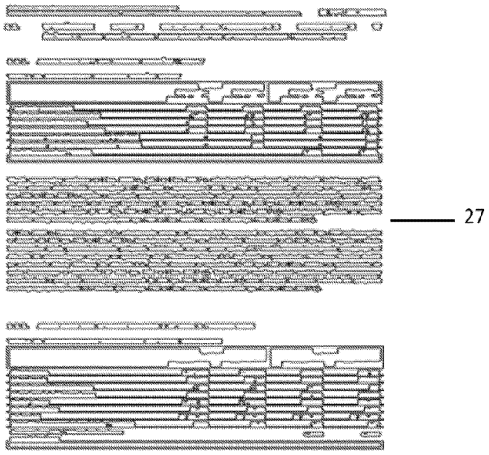
Фиг. 11



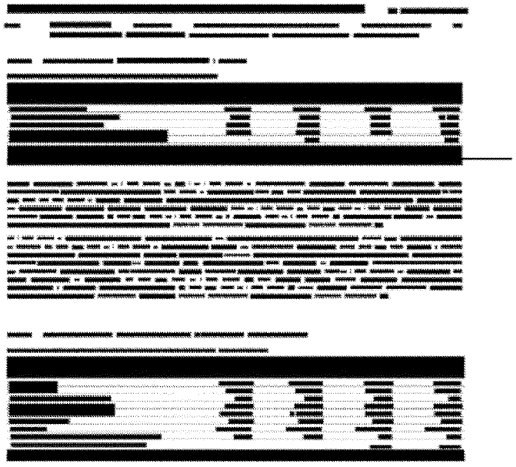
Фиг. 12



Фиг. 13



Фиг. 14



Фиг. 15