(19) Евразийское патентное ведомство

(11) 043496 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2023.05.29

(21) Номер заявки
201990647

(22) Дата подачи заявки
2019.04.03

(51) Int. Cl. *G06F 17/00* (2019.01)

(54) СПОСОБ И СИСТЕМА ДЛЯ ПРОВЕРКИ ЭЛЕКТРОННОГО КОМПЛЕКТА ДОКУМЕНТОВ

(31) **2019109055**
(32) **2019.03.28**
(33) **RU**
(43) **2020.09.30**

(71)(73) Заявитель и патентовладелец:
ПУБЛИЧНОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО "СБЕРБАНК РОССИИ" (ПАО СБЕРБАНК) (RU)

(72) Изобретатель:
Латышев Евгений Сергеевич, Тарасов Кирилл Геннадьевич (RU)

(74) Представитель:
Герасин Б.В. (RU)

(56) US-A1-20110255790
US-A1-20160354689
US-A1-20120281077

(57) В изобретении представленное техническое решение относится, в общем, к области анализа изображений, а в частности к способам и системам для проверки электронного комплекта документов, например отсканированных документов корпоративного клиента банка. Техническим результатом является повышение точности проведения автоматизированной проверки документов на их комплектность. Указанный технический результат достигается благодаря осуществлению способа проверки электронного комплекта документов, выполняемого по меньшей мере одним вычислительным устройством и содержащего этапы, на которых получают изображение документа, состоящего по меньшей мере из одной страницы; распознают символы на изображении страницы документа и преобразуют их в текстовую информацию; формируют вектор страницы документа на основе текстовой информации, полученной на предыдущем этапе; определяют на основе вектора страницы документа тип документа и тип его страницы; определяют перечень страниц и по меньшей мере один атрибут подписанта, наличие которых необходимо проверить в данном типе документа; проверяют наличие перечня страниц и по меньшей мере одного атрибута подписанта на полученном изображении документа для определения комплектности документа.

043496 B1

043496 B1

Область техники

Представленное техническое решение относится, в общем, к области анализа изображений, а в частности к способам и системам для проверки электронного комплекта документов, например, отсканированных документов корпоративного клиента банка.

Уровень техники

В настоящее время существует проблема оперативной и качественной обработки данных электронного комплекта отсканированных документов с целью проверки наличия обязательных для заранее определенных типов страниц атрибутов, таких как печать и/или подпись и/или комплект подписей. Из уровня техники известны различные решения, выполненные с возможностью обработки документов, например, клиента Банка, реализованные на базе ПО ABBYY InfoExtractor и пр. Также известно решение для проведения проверки комплекта документов, раскрытое в заявке US 2011134494 (A1), опубл. 09.06.2011 г., в котором осуществляют чтение документа, имеющего множество страниц; проверка данных изображения каждой страницы документа, имеющего множество страниц, при этом проверяются определенные области изображения документа на наличие в них информации и ее отсутствие. Данное решение является наиболее близким аналогом.

Существенным недостатком известных решений является отсутствие возможности проверить комплект отсканированных документов по следующим критериям:

комплектность пакета документов;

наличие печатей;

наличие и корректность состава подписей.

Раскрытие изобретения

Технической проблемой или задачей, поставленной в данном техническом решении, является создание нового эффективного метода автоматизированной проверки комплекта документов, например, документов корпоративного клиента Банка.

Техническим результатом является повышение точности проведения автоматизированной проверки документов на их комплектность. Дополнительным техническим результатом является повышение скорости проведения автоматизированной проверки документов на их комплектность.

Указанный технический результат достигается благодаря осуществлению способа проверки электронного комплекта документов, выполняемого по меньшей мере одним вычислительным устройством, и содержащего этапы, на которых:

получают изображение документа, состоящего из по меньшей мере одной страницы;

распознают символы на изображении страницы документа и преобразует их в текстовую информацию;

формируют вектор страницы документа на основе текстовой информации, полученной на предыдущем этапе;

определяют на основе вектора страницы документа тип документа и тип его страницы;

определяют перечень страниц и по меньшей мере один атрибут подписанта, наличие которых необходимо проверить в данном типе документа;

проверяют наличие перечня страниц и по меньшей мере одного атрибута подписанта на полученном изображении документа для определения комплектности документа.

В одном из частных примеров осуществления способа вектор страницы документа формируется на основе значений слов, содержащихся в текстовой информации, структуры зависимостей слов друг от друга и значений веса упомянутых слов.

В другом частном примере осуществления способа определение типа документа и типа его страницы на основе вектора страницы документа осуществляется посредством классификации документа по принадлежности к заранее определенным типам страниц и документов, причем математическая модель для классификации реализована посредством алгоритмов машинного обучения "случайный лес".

В другом частном примере осуществления способа этап проверки наличия по меньшей мере одного атрибута подписанта на полученном изображении документа, включает этапы, на которых

детектируют по меньшей мере один атрибут подписанта на изображении страницы документа для определения его расположения;

определяют, где атрибут подписанта должен находиться на данном типе страницы;

причем проверка наличия по меньшей мере одного атрибута подписанта на полученном изображении документа для определения комплектности документа осуществляется посредством сравнения информации о расположении атрибута подписанта на изображении страницы документа с информации, указывающей на то, где должен находиться атрибут подписанта на данном типе страницы.

В другом частном примере осуществления способа детектирование по меньшей мере одного атрибута подписанта осуществляется только на тех изображениях страниц документов, тип который указывает на то, что данные страницы содержат атрибуты подписанта.

В другом частном примере осуществления способа дополнительно классифицируют по меньшей мере один атрибут подписанта, причем классификация осуществляется на основе информации о расположения атрибута подписанта.

В другом частном примере осуществления способа атрибут подписанта представляют собой подпись и/или печать.

В другом предпочтительном варианте осуществления заявленного решения представлена система для проверки комплекта документов, содержащая по меньшей мере одно вычислительное устройство и по меньшей мере одну память, содержащую машиночитаемые инструкции, которые при их исполнении по меньшей мере одним вычислительным устройством выполняют вышеуказанный способ.

Краткое описание чертежей

Признаки и преимущества настоящего технического решения станут очевидными из приводимого ниже подробного описания изобретения и прилагаемых чертежей, на которых:

на фиг. 1 представлена общая схема взаимодействия элементов системы для проверки комплекта документов.

на фиг. 2 представлен пример отсканированного документа.

на фиг. 3 представлен пример общего вида системы для проверки комплекта документов.

Осуществление изобретения

Ниже будут описаны понятия и термины, необходимые для понимания данного технического решения.

В данном техническом решении под системой подразумевается, в том числе компьютерная система, ЭВМ (электронно-вычислительная машина), ЧПУ (числовое программное управление), ПЛК (программируемый логический контроллер), компьютеризированные системы управления и любые другие устройства, способные выполнять заданную, четко определенную последовательность операций (действий, инструкций).

Под устройством обработки команд подразумевается электронный блок либо интегральная схема (микропроцессор), исполняющая машинные инструкции (программы).

Устройство обработки команд считывает и выполняет машинные инструкции (программы) с одного или более устройств хранения данных. В роли устройства хранения данных могут выступать, но не ограничиваясь, жесткие диски (HDD), флеш-память, ПЗУ (постоянное запоминающее устройство), твердотельные накопители (SSD), оптические приводы.

Программа - последовательность инструкций, предназначенных для исполнения устройством управления вычислительной машины или устройством обработки команд.

База данных (БД) - совокупность данных, организованных в соответствии с концептуальной структурой, описывающей характеристики этих данных и взаимоотношения между ними, причем такое собрание данных, которое поддерживает одну или более областей применения (ISO/IEC 2382:2015, 2121423 "database").

В соответствии со схемой, приведенной на фиг. 1, система 1 для проверки комплекта документов содержит соединенные между собой: модуль 10 преобразования данных; модуль 20 классификации страниц, модуль 30 проверки атрибутов подписанта, таких как подписи и/или печати и модуль 40 проверки комплекта документов.

Указанные модули могут быть реализованы на базе программно-аппаратных средств системы 1 для проверки комплекта документов, например, на базе по меньшей мере одно вычислительного устройства, в частности микропроцессора, и по меньшей мере одной памяти, содержащей машиночитаемые инструкции для осуществления приписанных модулям ниже функций. Например, модуль 10 преобразования данных может содержать модуль 11 формирования векторов и модуль 12 фильтрации изображений, и может быть реализован на базе opensource-инструмента Tesseract (Tesseract Open Source OCR Engine) и алгоритма TF-IDF. Модуль 20 классификации страниц может быть реализован на базе заранее обученной математической модели с применением алгоритма обучения математической модели - случайный лес решающих деревьев (random forest). Модуль 30 проверки атрибутов подписанта может быть реализован на базе нейронной сети архитектуры YOLOv3, заранее обученной на типовом наборе подписей и печатей. Модуль 40 проверки комплекта документов может включать по меньшей мере одну БД 41 для хранения информации, которая может потребоваться для проверки комплекта документов.

В общем виде (см. фиг. 3) система (200) для проверки комплекта документов содержит объединенные общей шиной информационного обмена один или несколько процессоров (201), средства памяти, такие как ОЗУ (202) и ПЗУ (203), интерфейсы ввода/вывода (204), устройства ввода/вывода (205), и устройство для сетевого взаимодействия (206).

Процессор (201) (или несколько процессоров, многоядерный процессор и т.п.) может выбираться из ассортимента устройств, широко применяемых в настоящее время, например, таких производителей, как: Intel™, AMD™, Apple™, Samsung Exynos™, MediaTEK™, Qualcomm Snapdragon™ и т.п. Под процессором или одним из используемых процессоров в системе (200) также необходимо учитывать графический процессор, например, GPU NVIDIA с программной моделью, совместимой с CUDA, или Graphcore, тип которых также является пригодным для полного или частичного выполнения способа, а также может применяться для обучения и применения моделей машинного обучения в различных информационных системах.

ОЗУ (202) представляет собой оперативную память и предназначено для хранения исполняемых процессором (201) машиночитаемых инструкций для выполнения необходимых операций по логической

обработке данных. ОЗУ (202), как правило, содержит исполняемые инструкции операционной системы и соответствующих программных компонент (приложения, программные модули и т.п.). При этом, в качестве ОЗУ (202) может выступать доступный объем памяти графической карты или графического процессора.

ПЗУ (203) представляет собой одно или более устройств постоянного хранения данных, например, жесткий диск (HDD), твердотельный накопитель данных (SSD), флэш-память (EEPROM, NAND и т.п.), оптические носители информации (CD-R/RW, DVD-R/RW, BlueRay Disc, MD) и др.

Для организации работы компонентов системы (200) и организации работы внешних подключаемых устройств применяются различные виды интерфейсов В/В (204). Выбор соответствующих интерфейсов зависит от конкретного исполнения вычислительного устройства, которые могут представлять собой, не ограничиваясь: PCI, AGP, PS/2, IrDa, FireWire, LPT, COM, SATA, IDE, Lightning, USB (2.0, 3.0, 3.1, micro, mini, type C), TRS/Audio jack (2.5, 3.5, 6.35), HDMI, DVI, VGA, Display Port, RJ45, RS232 и т.п.

Для обеспечения взаимодействия пользователя с вычислительной системой (200) применяются различные средства (205) В/В информации, например, клавиатура, дисплей (монитор), сенсорный дисплей, тач-пад, джойстик, манипулятор мышь, световое перо, стилус, сенсорная панель, трекбол, динамики, микрофон, средства дополненной реальности, оптические сенсоры, планшет, световые индикаторы, проектор, камера, средства биометрической идентификации (сканер сетчатки глаза, сканер отпечатков пальцев, модуль распознавания голоса) и т.п.

Средство сетевого взаимодействия (206) обеспечивает передачу данных посредством внутренней или внешней вычислительной сети, например, Интранет, Интернет, ЛВС и т.п. В качестве одного или более средств (206) может использоваться, но не ограничиваться: Ethernet карта, GSM модем, GPRS модем, LTE модем, 5G модем, модуль спутниковой связи, NFC модуль, Bluetooth и/или BLE модуль, Wi-Fi модуль и др.

Дополнительно могут применяться также средства спутниковой навигации в составе системы (200), например, GPS, ГЛОНАСС, BeiDou, Galileo. Конкретный выбор элементов устройства (200) для реализации различных программно-аппаратных архитектурных решений может варьироваться с сохранением обеспечиваемого требуемого функционала.

На первом этапе работы системы 1 на модуль 10 преобразования данных поступает по меньшей мере одно изображение документа, в частности отсканированного документа, например, файл в формате многостраничного PDF, JPEG, TIFF или любого другого известного формата, который может использоваться для хранения в нем отсканированного электронного комплекта документа. Изображение документа может поступать от источника данных изображений 50, в частности непосредственно от устройства сканирования документов, например, сканера, либо могут быть извлечены из соответствующей базы данных изображений, в которую данные изображения документов заранее сохранены.

Документом, изображение которого поступает на модуль 10 преобразования данных, может быть любой документ, состоящий по меньшей мере из одной страницы, которая может содержать атрибуты подписанта, и заполненный в соответствии с известным шаблоном. Документ может быть, например, договором, заключенным между компаниями "А" и "Б", либо между компанией и физическим лицом, либо между физическими лицами, либо документ может представлять такой вид документа, который подписывается только лишь одним подписантом - компанией или физическим лицом, например, доверенностью от компании или от физического лица; или пр. Модуль 10 преобразования данных осуществляет распознавание символов на по меньшей мере одном изображении страницы документа и преобразует их в текстовую информацию. Также модуль 10 преобразования данных может быть выполнен с возможностью предобработки полученной текстовой информации для снижения многообразия возможных текстов распознанных изображений документов с целью упростить работу следующим модулям системы. На первом этапе осуществляется токенизация текстовой информации. Этап токенизации, предполагает выделение базовых элементов текста (токенов), ограниченных с двух сторон разделительными символами, пробелами или знаками пунктуации. Элементами здесь выступают слова, числа, даты, сокращения, аббревиатуры, составные предлоги и т.д. Токенизация позволяет выделить дискретные единицы текста, являющиеся основой для дальнейшей работы на этапах морфологического и синтаксического анализа. В результате токенизации каждому элементу присваивается соответствующий тип: слово, число, дата, адрес и т.д.

Далее модуль 10 преобразования данных переходит к этапу формирования векторов страницы документа посредством модуля 11 формирования векторов. На данном этапе упомянутый модуль для каждого слова, полученного после обработки текста, определяет значение веса слова с помощью статистической меры TF-IDF.

TF-IDF - статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса.

Вес некоторого слова пропорционален количеству употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции.

Мера TF-IDF часто используется в задачах анализа текстов и информационного поиска, например, как один из критериев релевантности документа поисковому запросу, при расчёте меры близости документов при кластеризации.

TF (term frequency - частота слова) - отношение числа вхождения некоторого слова к общему количеству слов документа. Значимость слова в пределах отдельного документа может быть определена следующей характеристикой:

$$tf(t,d) = \frac{n_i}{\sum_k n_k},$$

где n_i - число вхождений слова t_i в документ d;

$$\sum_k n_k$$

общее число слов в данном пользовательском запросе и/или документе.

IDF (inverse document frequency - обратная частота документа) - величина, обратно пропорциональная частоте, с которой некоторое слово встречается в документах коллекции.

Учёт IDF уменьшает вес широкоупотребительных слов. Для каждого уникального слова в пределах конкретной коллекции документов существует только одно значение IDF. IDF-характеристика определяется следующим отношением:

$$idf\ (t,D) = log\frac{|D|}{|d_i \supset t_i|}$$

где

$$|D|$$

количество документов в корпусе;

$$|d_i \supset t_i|$$

количество документов, в которых встречается t_i.

Таким образом, мера TF-IDF является произведением двух сомножителей

$$tf \cdot idf(t,d,D) = tf(t,d) \times idf(t,D).$$

Большой вес в мере TF-IDF получают слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

Мера TF-IDF часто используется для представления документов коллекции в виде числовых векторов, отражающих важность использования каждого слова из некоторого набора слов (количество слов набора определяет размерность вектора) в каждом документе. Подобная модель называется векторной моделью и даёт возможность сравнивать тексты, сравнивая представляющие их вектора в какой-либо метрике (евклидово расстояние, косинусная мера, манхэттенское расстояние, расстояние Чебышева и др.), то есть, производя кластерный анализ.

Далее модуль 11 формирования векторов на основе значений слов, полученных после предобработки, структуры зависимостей слов друг от друга в текстовой информации и значений веса упомянутых слов формирует вектор страницы документа. Сформированный вектор страницы документа направляются в модуль 20 классификации страниц для определения типа документа и типа страниц документа, т.е. для классификации документа по принадлежности к заранее определенным типам страниц и документов.

Соответственно, если отсканированный документ содержит две и более страницы, то модуль 11 формирования векторов формирует аналогичным образом для каждой страницы документа вектор страницы документа, которые также направляются в модуль 20 классификации страниц.

Модуль 20 классификации страниц для определения типа документа и типов его страниц содержит математическую модель, на вход которой поступают данные о векторах страниц документа. Математическая модель может быть реализована посредством алгоритмов машинного обучения "случайный лес" (Random forest), заключающихся в использовании комитета (ансамбля) решающих деревьев. Классификация объектов проводится путём голосования: каждое дерево комитета относит классифицируемый объект, в данном случае страницу документа, к одному из классов, характеризующий тип страницы и тип документа, и побеждает класс, за который проголосовало наибольшее число деревьев. Оптимальное число деревьев подбирается таким образом, чтобы минимизировать ошибку классификатора на тестовой выборке.

Соответственно, после обработки данных о векторе страницы документа на выход математической модели поступает от каждого дерева решений указатели типа документа и типа страницы. Модуль 20 классификации страниц анализирует количество упомянутых указателей, полученных на выходе упомянутой модели, и определяет тип документа и тип страницы на основе того указателя типа документа, количество которых на выходе математической модели больше, т.е. за который проголосовало наибольшее число деревьев. Например, если наибольшее число деревьев проголосовало за то, что документ, на основе изображения страницы которого был сформирован вектор страницы документа, является Договором, заключенным между компаниями "А" и "Б", а страница является страницей Договора с атрибутами подписантов, то тип документа будет определяться как "договор", а данные о типе страницы будут указывать на то, что страница на изображении является страницей договора, которая должна содержать атрибуты подписантов в виде подписей и печатей компаний "А" и "Б", расположенных в заданных областях на странице (например, в областях 105 и 106, см. фиг. 2). Также тип документа может быть определен, например, как доверенность от компании "А" или от физического лица, а тип страницы - страница доверенности с атрибутами подписанта, например, в виде подписи и печати в заданной области страни-

цы, например, области 106, если доверенность от компании, или только в виде подписи в заданной области страницы, если доверенность от физического лица.

Если отсканированный документ состоит из двух и более страниц, то на вход математической модели поступают данные о двух и более векторах страниц документа. Модуль 20 классификации страниц аналогичным образом анализирует количество упомянутых указателей на выходе математической модели и определяет тип документа и перечень его страниц на основе того указателя типа документа, количество которых на выходе математической модели больше. В данном случае тип документа определяется на основе векторов всех его страниц. Например, на основе векторов страниц документа модуль 20 классификации страниц может определить тип документа как договор между компаниями "А" и "Б", состоящий из 4 станицы, причем данные о типах страниц могут указывать на то, что первая страница является страницей Договора, не содержащей атрибутов подписанта, вторая страница - страница договора с атрибутами подписантов в заданных областях, а 3 и 4 страницы - являются приложениями, не содержащими атрибуты подписантов.

Данные о типах документа и типах его страниц модуль 20 классификации страниц направляет в модуль 40 проверки комплекта документов и в модуль 12 фильтрации изображений, который определяет типы страниц с по меньшей мере одним атрибутом подписанта, извлекает соответствующие изображения страниц с по меньшей мере одним атрибутом подписанта из изображения документа и направляет данные изображения страниц в модуль 30 проверки атрибутов подписанта для дальнейшего анализа. Таким образом, поскольку в модуль 30 проверки атрибутов подписанта направляется не все изображение документа, а только изображения страниц документа, тип которых предполагает наличие на данных страницах по меньшей мере одного атрибута подписанта, снижается вычислительная нагрузка и повышается скорость обработки изображений модулем 30 для детектирования изображений атрибутов подписанта, вследствие чего повышается скорость проведения автоматизированной проверки документов на их комплектность.

Модуль 30 проверки атрибутов подписанта после получения изображений страниц с по меньшей мере одним атрибутом подписанта переходит к этапу детектирования на каждом полученном изображении страницы документа по меньшей мере одного изображения атрибута подписанта для определения его расположение на странице документа. Например, модуль 30 проверки атрибутов подписанта может определить, что изображение атрибута подписанта представляет собой изображение подписи и/или печати в области 105 или 106 документа (см. фиг 2). Соответственно, в области 101 документа 100 может содержаться информация о номере Договора, в области 102 - название города, в области 103 - дата Договора, а в области 104 - текст Договора. Для детектирования изображений атрибутов подписанта используются известные алгоритмы работы нейронной сети архитектуры YOLOv3, обученной на отобранном наборе данных подписей и печатей, раскрытые, например, в статье, опубликованной в Интернет по адресу: https://pireddie.com/media/files/papers/YOLOv3.pdf. Данные о детектированных атрибутах подписанта, в частности информация об их расположении на странице документа, передаются в модуль 40 проверки комплекта документов. Модуль 40 проверки комплекта документов в процессе своей работы проверяет наличие обязательных для данного типа документа перечня страниц и атрибутов подписантов, таких как печать и/или подпись и/или комплект подписей, в заданных областях страниц. Для определения атрибутов подписанта, наличие которых необходимо проверить, модуль 40 проверки комплекта документов может быть оснащен соответствующей БД 41 с информацией о шаблонах документов, их перечня страниц, и атрибутах подписантов, наличие которых необходимо проверить в заданной области страниц из перечня страниц данного типа документа. Поскольку составление векторов страниц осуществляется на основе текстовой информации, которая может включать названия одной или нескольких компаний, или имена одного или нескольких физических лиц, то информация о типе страниц также будет определять, в какой области страницы должны располагаться атрибуты подписана на изображении страницы документа. Например, если информация о типе документа указывает на то, что данный документ является Договором 200 (см. фиг. 2), состоящим из 1 страницы Договора, который должен быть подписан только лишь одним подписантом, то модуль 40 проверки комплекта документов в соответствии с шаблоном документа проверяет область страницы 105 или 106, в зависимости от типа документа и типа страницы, на наличие атрибута подписанта, в частности его подписи и/или печати. Если информация о типе документа указывает на то, что данный документ является договором, состоящим из 2 страниц договора, причем вторая страница Договора в соответствии с шаблоном документа должна быть подписана двумя подписантами, то модуль 40 проверки комплекта документов проверяет области 105 или 106 второй страницы на наличие атрибутов подписантов, причем расположение атрибутов первого и второго подписантов в упомянутых областях определяется типом документа и типами его страниц.

Для проверки документа модуль 40 проверки комплекта документов на основе данных о типе документа, полученных от модуля 20 классификации страниц, осуществляет поиск в БД 41 шаблона данного типа документа, на основе которого модуль 40 будет выполнять проверку комплекта документа, и извлекает информацию о типах страниц данного шаблона документа. Например, если модуль 20 классификации страниц определил, что отсканированный документ является Договором между компаниями "А" и "Б", то на основе данной информации о типе документа модуль 40 проверки комплекта документов на-

ходит в БД шаблон Договора между компаниями "А" и "Б" и извлекает информацию о типах страниц, присутствующих в шаблоне Договора. Если Договор выполнен на 1 листе, то как правило атрибуты подписанта должны быть расположены на первой странице документа. Информация о том, что атрибуты подписанта должны находиться на первой странице, а также их расположение на странице, может содержаться в информации о типе страницы, в соответствии с которой модуль 40 проверки комплекта документов будет осуществлять проверку наличия атрибутов подписантов на первой странице договора.

Если Договор состоит двух и более страниц, то, например, информация о типе последней страницы документа может содержать информацию о том, что атрибуты подписанта должны находиться в заданной области (например, в областях 105 или 106) на данной странице. Также информация о типе первой страницы или о типе документа может содержать информацию о том, что атрибуты подписанта должны находиться в заданных областях на второй или другой странице в документе.

Если информация о типе страниц, полученной от модуля 20 классификации страниц, не совпадают с информацией о типе станиц шаблона документа, то модуль 40 проверки комплекта документов принимает решение о том, что комплект документа неполон. Например, согласно шаблону документа данный отсканированный документ является Договором, заключенным с физическим лицом, состоящим из 3 страниц, где первые 2 страницы являются страницами Договора, а третья страница - сканом паспорта. Таким образом, если в отсканированном Договоре будет отсутствовать скан паспорта или вместо скана паспорта будет приложен другой документ, изображение которого будет обработано системой 1, то информация о типе третьей странице, полученная от модуля 20, не будет совпадать с информацией о типе станиц шаблона документа. Информация о том, что отсканированный комплект документа неполон, например, в виде сообщения "отсутствует скан паспорта", может быть выведена на средства (205) В/В информации.

Если информация о типе страниц, полученной от модуля 20 классификации страниц, совпадают с информацией о типе станиц шаблона документа, то модуль 40 проверки комплекта документов извлекает из БД 41 информацию о расположении по меньшей мере одного атрибута подписанта на по меньшей мере одной странице согласно шаблону документа для тех типов страниц, которые должны содержать по меньшей мере один атрибут подписанта. Упомянутая информация о расположении по меньшей мере одного атрибута подписанта может быть получена экспериментально на основе данных о средних координатах расположения подписей и печатей в шаблонах документов. Например, если отсканированный документ является Договором между компаниями "А" и "Б", состоящий из 1 страницы, то модуль 20 классификации данных извлекает из БД 41 информацию о расположении подписей и/или печатях (т.е. о атрибутах подписанта) на данной странице документа в соответствии с шаблоном. В частности, в БД 41 может храниться как тип шаблона документа, в котором информация о расположении будет указывать на то, что атрибуты подписанта компании "А" должны находиться в области страницы 105 Договора 100, а атрибуты подписанта компании "Б" - в области 106 Договора 100, так и тип шаблона документа, в котором атрибуты подписанта компании "Б" должны находиться в области 105 Договора 100, а атрибуты подписанта компании "А" - в области 106 Договора 100.

Соответственно, извлеченную на предыдущем шаге из БД 41 информацию о расположении по меньшей мере одного атрибута подписанта модуль 40 проверки комплекта документов сравнивает с информацией о расположении по меньшей мере одного атрибута на странице документа, полученной от модуля 30. Если извлеченная из БД 41 упомянутая информация о расположении по меньшей мере одного атрибута подписанта не совпадает с информацией о расположении по меньшей мере одного атрибута на странице документа, полученной от модуля 30, то модуль 40 проверки комплекта документов принимает решение о том, что комплект документа неполон. Информация о том, что отсканированный комплект документа неполон, например, в виде сообщения "отсутствует подпись клиента на 3 странице", может быть выведена на средства (205) В/В информации Если извлеченная из БД 41 упомянутая информация о расположении совпадает с информацией о расположении, полученной от модуля 30, то модуль 40 проверки комплекта документов принимает решение о том, что комплект документа соответствует установленным требованиям комплектности. Информация о том, что отсканированный комплект документа полон также может быть выведена на средства (205) В/В информации.

Таким образом, за счет того, что тип документа и типы страниц документа, входящих в его состав, определяются на основе векторов страниц, сформированных на основе текстовой информации, содержащейся на странице, повышается точность определения типа документа и перечня его станиц, а также перечня страниц и атрибутов подписанта, наличие которых необходимо проверить на изображении документа, т.е. обеспечивается повышение точности проведения автоматизированной проверки документов на их комплектность. Формирование векторов страниц с учетом значений слов, содержащихся в текстовой информации, структуры зависимостей слов друг от друга и значений веса упомянутых слов дополнительно повысит точность при определении типа документа и типов страниц документа, а также перечня страниц и атрибутов подписанта, наличие которых необходимо проверить на изображении документа. Дополнительно модуль 40 проверки комплекта документов может быть выполнен с возможностью классификации по меньшей мере одного атрибута подписанта, которая осуществляется на основе информации о расположения атрибута подписанта. Для обеспечения данной возможности БД 41 дополнительно

содержит информацию о том, к какой стороне Договора относится атрибут подписанта в зависимости от его расположения на странице. Например, в БД 41 может содержаться информация о том, что в области страницы 105 Договора 100 расположен атрибут подписанта клиента, а области страницы 106 - исполнителя Договора. Таким образом, сравнивая информацию о расположении атрибута подписанта, полученную от модуля 30, с информацией о расположении атрибута подписанта из БД 41 модуль 40 проверки комплекта документов классифицирует изображение атрибута подписанта, например, как атрибут подписанта клиент, если атрибут подписанта расположен в области страницы 105, или как атрибут подписанта исполнителя Договора, если атрибут подписанта расположен в области страницы 106.

Модификации и улучшения вышеописанных вариантов осуществления настоящего технического решения будут ясны специалистам в данной области техники. Предшествующее описание представлено только в качестве примера и не несет никаких ограничений. Таким образом, объем настоящего технического решения ограничен только объемом прилагаемой формулы изобретения.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ проверки электронного комплекта документов, выполняемый по меньшей мере одним вычислительным устройством, содержащий этапы, на которых

получают изображение по меньшей мере одного документа, состоящего по меньшей мере из одной страницы, входящей в упомянутый комплект документов;

распознают символы на изображении страницы документа и преобразуют их в текстовую информацию;

формируют вектор страницы документа на основе текстовой информации, полученной на предыдущем этапе;

определяют на основе вектора страницы документа тип документа и тип его страницы;

определяют перечень страниц и по меньшей мере один атрибут подписанта, который должен присутствовать по меньшей мере на одной странице определенного типа, наличие которых необходимо проверить в данном типе документа;

проверяют наличие страниц, соответствующих упомянутому перечню; и по меньшей мере одного атрибута подписанта для проверяемого типа документа, который должен присутствовать по меньшей мере на одной странице определенного типа, в комплекте документов.

2. Способ по п.1, отличающийся тем, что вектор страницы документа формируется на основе значений слов, содержащихся в текстовой информации, структуры зависимостей слов друг от друга и значений веса упомянутых слов.

3. Способ по п.1, отличающийся тем, что определение типа документа и типа его страницы на основе вектора страницы документа осуществляется посредством классификации документа по принадлежности к заранее определенным типам страниц и документов, причем математическая модель для классификации реализована посредством алгоритмов машинного обучения "случайный лес".

4. Способ по п.1, отличающийся тем, что этап проверки наличия по меньшей мере одного атрибута подписанта на полученном изображении документа включает этапы, на которых

детектируют по меньшей мере один атрибут подписанта на изображении страницы документа для определения его расположения;

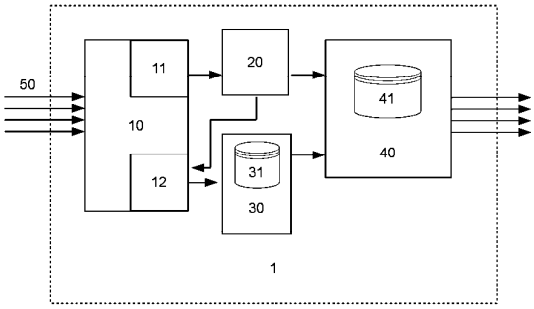
определяют, где атрибут подписанта должен находиться на данном типе страницы;

причем проверка наличия по меньшей мере одного атрибута подписанта на полученном изображении документа для определения комплектности документа осуществляется посредством сравнения информации о расположении атрибута подписанта на изображении страницы документа с информации, указывающей на то, где должен находиться атрибут подписанта на данном типе страницы.

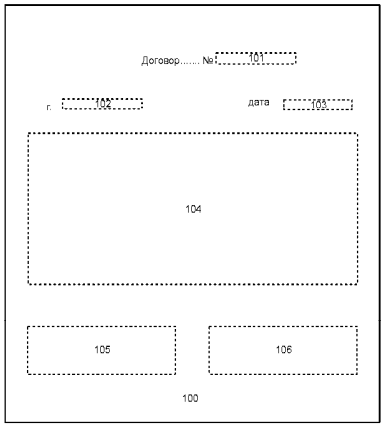
5. Способ по п.4, отличающийся тем, что детектирование по меньшей мере одного атрибута подписанта осуществляется только на тех изображениях страниц документов, тип который указывает на то, что данные страницы содержат атрибуты подписанта.

6. Способ по п.1, отличающийся тем, что атрибут подписанта представляют собой подпись и/или печать.

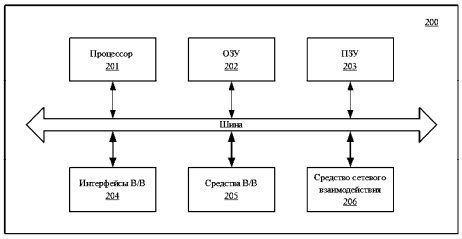
7. Система для проверки комплекта документов, содержащая по меньшей мере одно вычислительное устройство и по меньшей мере одну память, содержащую машиночитаемые инструкции, которые при их исполнении по меньшей мере одним вычислительным устройством выполняют способ по любому из пп.1-6.



Фиг. 1



Фиг. 2



Фиг. 3