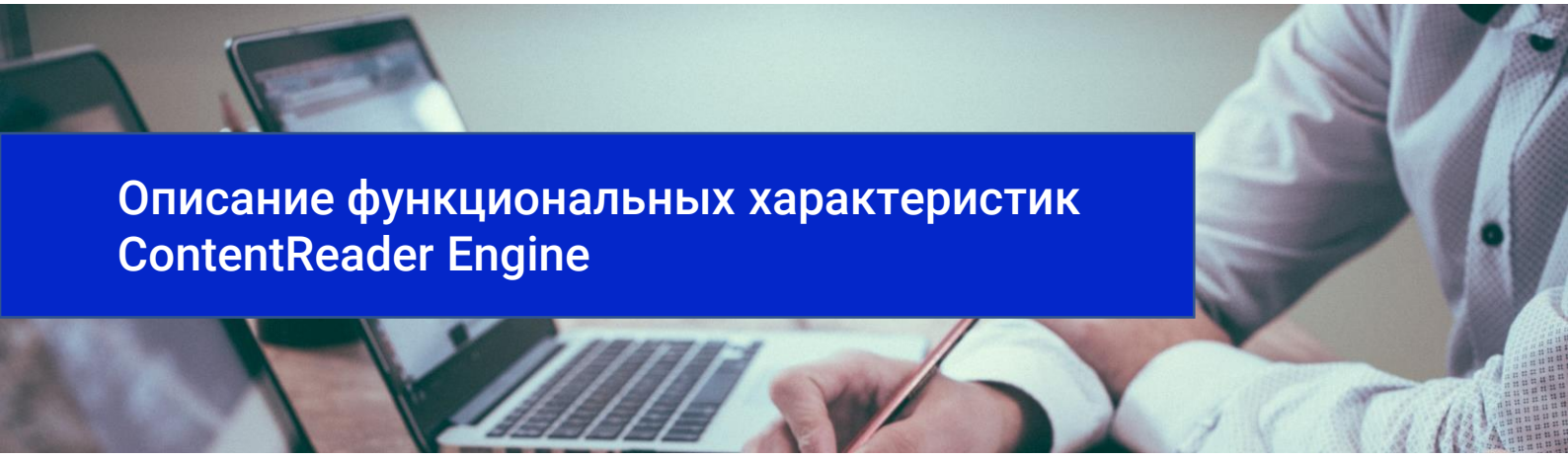


Описание функциональных характеристик ContentReader Engine

Технология распознавания печатного текста (OCR)

Технология доступна для более чем 200 языков:

- Европейские языки: латиница, кириллица, армянский и греческий алфавиты
- Другие языки: китайский, японский, корейский, арабский, фарси, тайский, вьетнамский, иврит, бирманский
- Распознавание шрифтов OCR-A, OCR-B, MICR (E13B) и CMC7 и документов, напечатанных на матричных принтерах или пишущих машинках

Технология распознавания рукопечатного текста (ICR)

Технология доступна для более чем 120 языков:

- Европейские и другие языки
- 22 региональных рукопечатных стиля
- Распознавание рукопечатных символов в полях и рамках
- Распознавание индийских цифр, используемых в арабских государствах

Возможно распознавание рукопечатной информации на разных языках одновременно (многоязычный ICR).

Технология распознавания штрихкодов (OBR)

- Поддержка одномерных и двухмерных штрихкодов
- Автоматическое определение и распознавание штрихкодов, расположенных на документе под любым углом

Инструменты конвертации документов в PDF и PDF/A

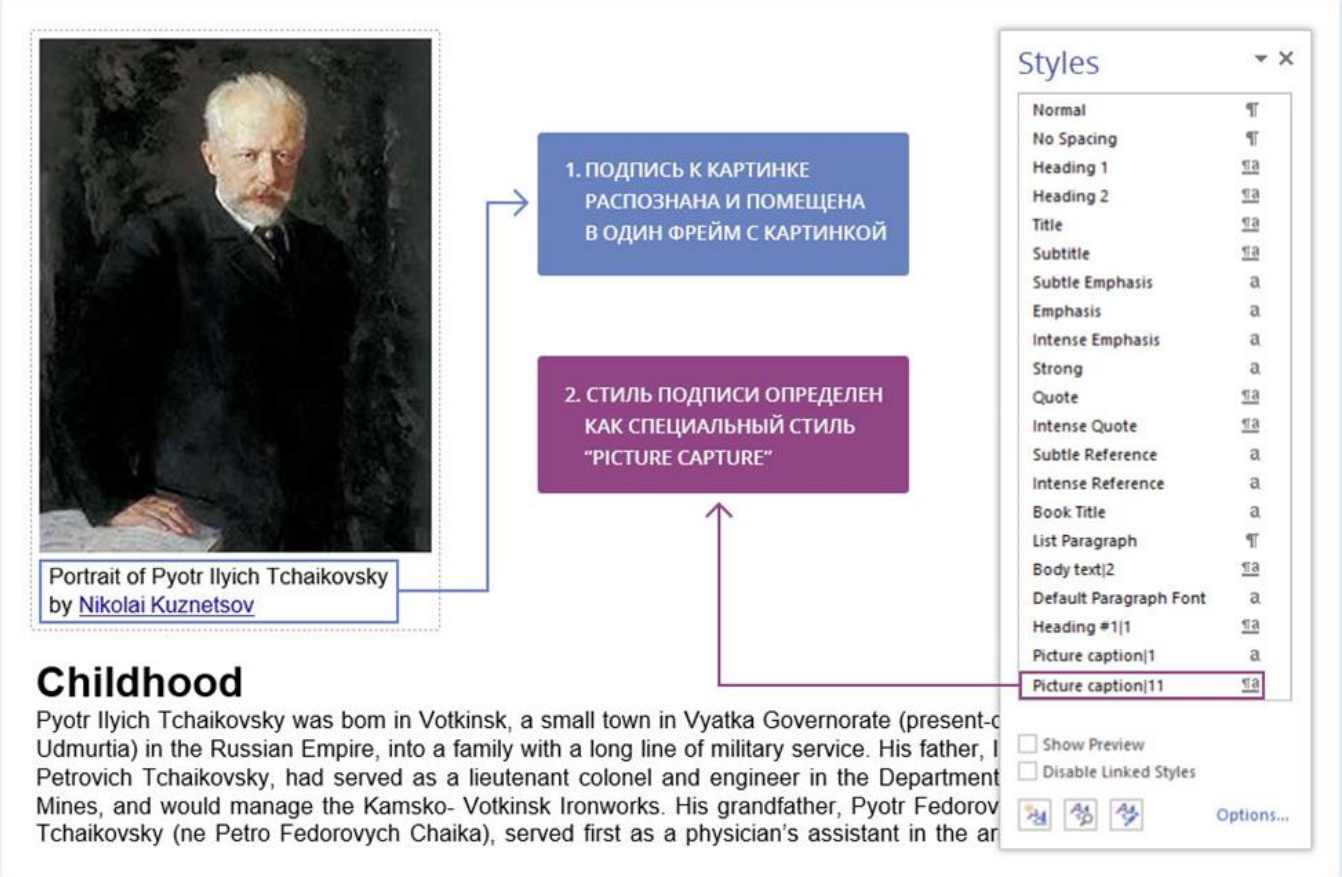
OCR SDK позволяет конвертировать сканы, цифровые фотографии, TIFF, JPEG, BMP и другие файлы различных форматов во множество форматов PDF и PDF/A с возможностью поиска. Кроме того, можно импортировать файлы PDF и PDF/A и обрабатывать их разными способами.

Конвертируйте документы в PDF или обрабатываете файлы PDF и PDF/A — ContentReader Engine позволяет обрабатывать и создавать электронные документы в соответствии со стандартами PDF/A-3 и электронные счета ZUGFeRD с помощью целого ряда возможностей и инструментов.

Технология определения и восстановления логической структуры и форматирования документа

Для анализа оформления документов и оценки гипотез ContentReader Engine использует алгоритмы на базе искусственного интеллекта, машинного обучения и адаптивной технологии распознавания документов Adaptive Document Recognition Technology (ADRT). На этапе анализа документа программа разбивает его на отдельные страницы и проверяет оформление каждой из них, а именно, где расположен текст, изображения, штрихкоды и элементы таблиц. Параллельно проверяется логическая структура документа. Таким образом инструмент определяет роль текстовых элементов — например, колонтитулы определяются как колонтитулы, а не как фрагменты отдельных страниц.

Информация о тексте, изображениях и элементах форматирования сохраняется и используется на этапе итогового воссоздания документа. В результате получается точно воссозданный документ, например, в формате Word, с сохранением элементов форматирования, таких как таблицы, колонтитулы, номера страниц, сноски, содержание и многое другое.



Воссоздание документа: логическая структура, элементы и форматирование

- Иерархическая структура заголовков
- Заголовки для изображений/таблиц/диаграмм
- Содержание
- Верхние и нижние колонтитулы
- Шрифты и стили шрифтов
- Нумерация страниц
- Сноски
- Логическая последовательность текста
- Воссоздание пунктов маркированного списка и нумерации
- Сохранение гиперссылок

Параллельная обработка многостраничных и одностраничных документов

Гибкая и масштабируемая архитектура ContentReader Engine позволяет использовать многоядерные процессоры для обработки изображений в параллельных потоках, что значительно повышает скорость распознавания.

По умолчанию ContentReader Engine определяет, использовать ли параллельную обработку автоматически в зависимости от нескольких факторов, таких как, число доступных физических и логических ядер процессора, число ядер в параметрах лицензии и числа страниц в документе. При необходимости настройки параллельной обработки можно изменить и выбрать необходимое число активных процессов.

ContentReader Engine поддерживает два разных объекта, за счет которых обеспечивается параллельная обработка — CRDocument и BatchProcessor. Выбирайте объекты в зависимости от сценария использования.

Конвертация многостраничных документов с большим числом страниц

Данный сценарий подразумевает обработку больших документов и книг. В этом случае можно распараллелить распознавание страниц документа, а затем выполнить этапы синтеза и экспорта в главном процессе. Кроме того, можно организовать одновременную обработку нескольких многостраничных документов с использованием нескольких экземпляров SDK (pool of Engines), но при этом нужно быть готовым к тому, что возможны существенные утечки памяти, которые могут привести к сбоям программы.

Для параллельной обработки многостраничных документов Контент ИИ рекомендует использовать объект CRDocument. Использование данного объекта наименее трудозатратно, поскольку разработчику не нужно реализовывать никаких дополнительных интерфейсов. В этом случае параллельно производятся только предобработка, анализ и распознавание, а синтез и экспорт выполняются последовательно в главном процессе.

Конвертация большого числа одностраничных документов

Используется, когда необходимо обрабатывать большое количество одностраничных документов. В этом случае применить распараллеливание легко, поскольку все страницы не зависят друг от друга и не требуют одновременного выделения большого количества памяти.

ContentReader Engine позволяет решить данную задачу двумя способами:

Использование BatchProcessor

Преимущества этого метода заключаются в том, что его можно использовать для любого числа и типа документов, которые необходимо обработать прямо при поступлении. Этот метод, однако, требует больше усилий: пользователям понадобится реализовать интерфейсы файлового адаптера и пользовательского источника изображений. Документы открываются, предварительно обрабатываются, анализируются и распознаются параллельно.

Использование нескольких экземпляров SDK (a pool of Engines), загруженных в качестве out-of-process COM-сервера

Это наиболее эффективный метод, который позволяет обеспечить высокую скорость и автоматически устраняет трудности, связанные с использованием многопоточности: все операции с объектами ContentReader Engine сериализуются при помощи COM.

Инструменты для предварительной обработки изображения документа

После получения изображений ContentReader Engine выполняет их предварительную обработку, что позволяет улучшить качество документа и оптимизировать процесс распознавания данных. Таким образом, даже изображения самого низкого качества и документы, сфотографированные на смартфон, эффективно обрабатываются и распознаются в максимально высоком качестве.

Базовые функции обработки изображений

ContentReader Engine позволяет выполнять следующие действия с изображениями, например:

- Изменять масштаб
- Обрезать изображение
- Делать обтравку изображения
- Создавать изображения для предпросмотра
- Поворачивать изображение (на 90, 180 и 270 градусов)
- Выпрямлять текстовые строки
- Создавать зеркальное отражение и инвертирование
- Удалять шумы
- Повышать контрастность

Продвинутые функции обработки изображений

- Технология Camera-OCR
- Предварительная обработка документов с печатями и рукописными комментариями
- Автоматическое разделение двойных страниц
- Автоматическое определение ориентации страниц (90, 180 и 270 градусов)
- Автоматическое выравнивание изображений (до +/- 20 градусов)
- Удаление пятен (очистка изображения)
- Очистка изображений в отдельных блоках
- Фильтрация текстур и адаптивная бинаризация
- Редактирование текста и цвета фона
- Распознавание информации из полей с разными границами и рамками

Непревзойденное качество обработки

Документы, сфотографированные на цифровые камеры, телефоны и планшеты зачастую обладают высоким качеством, однако, в зависимости от устройства, им свойственны некоторые искажения. Интеллектуальная технология позволяет определять фотографии, сделанные на цифровую камеру, и активировать алгоритмы обработки таких изображений, чтобы устранять искажения, размытость, искривление текстовых строк, отсутствие информации о разрешении или ошибки, возникшие из-за недостаточного освещения.

Корректировка искажений перспективы

Искажения перспективы вызывают разные трудности при распознавании текста:

- Потенциальные ошибки распознавания символов
- Ошибки при разделении страниц
- Изменения размера шрифта (сверху вниз)

Корректировка размытых изображений

При использовании камеры без штатива можно получить размытое изображение. Этот дефект, не заметный на экране камеры, может приводить к ошибкам распознавания. После обработки полученного изображения бинарное изображение выглядит «читаемым»

Уменьшение цифорового шума (ISO)

Цифровой шум выглядит на изображении как множество маленьких пикселей разных цветов. Этот дефект изображения приводит к ошибкам бинаризации и потере символов. Благодаря специальному фильтру ContentReader Engine уменьшает цифровой шум и выравнивает фон, предотвращая потерю информации:

Готовые решения для распознавания визитных карточек и MRZ

Готовое решение для распознавания машинно-считываемой зоны (MRZ) в документах, удостоверяющих личность

Во многих документах, удостоверяющих личность, персональная информация кодируется в машиночитаемых зонах, как указано ICAO в Doc 9303. Готовая к использованию технология позволяет автоматически извлекать полевые данные из машинных зон в документах, удостоверяющих личность, и проверять соответствующие контрольные цифры. Результаты могут быть экспортированы в формате XML или JSON. Функция добавляет значительную ценность системам для быстрого захвата и верификации персональных данных - например, во время открытия счета в банке или при верификации клиентов.

Распознавание визитных карточек

В ContentReader Engine интегрирована готовая к использованию технология для распознавания визитных карточек. API обладает широким спектром функций распознавания таких документов, от предварительной обработки до обеспечения доступа к распознанной информации.

Список распознаваемых полей:

- ФИО;
- Название компании;
- Должность;
- Адрес;
- Адрес электронной почты;
- Номера телефонов;
- Факс;
- Номера мобильного телефона;
- Веб-сайт

Экспорт в формат vCard

Распознанные данные могут быть сохранены в формате vCard, который используется для передачи контактных данных по электронной почте.

Авторазделение визитных карточек, отсканированных на одной странице

Обычно пользователи фотографируют или сканируют сразу несколько визитных карточек. Технология поддерживает работу с несколькими визитками, отсканированными или сфотографированными на одном листе. Такие страницы предварительно разбиваются на несколько страниц по одной визитке на каждой и затем передаются в дальнейшую обработку.

Распознавание визитных карточек на 27 языках

- Английский
- Венгерский
- Греческий
- Датский
- Индонезийский
- Итальянский
- Испанский
- Китайский (традиционный)
- Китайский (упрощенный)
- Корейский
- Немецкий
- Нидерландский (Нидерланды)
- Норвежский
- Норвежский (Нюнорск)
- Норвежский (Букмол)
- Польский
- Португальский (Бразилия)
- Португальский (Португалия)
- Русский
- Турецкий
- Украинский
- Финский
- Французский
- Чешский
- Шведский
- Эстонский
- Японский

Инструмент для обучения графического и/или текстового классификатора

Инструментарий ContentReader Engine включает технологию для классификации документов, что позволяет создавать приложения для автоматического распределения документов по предопределенным категориям и классам. В передовых алгоритмах классификации используются технологии машинного обучения и обработки естественного языка, которые позволяют выявить малейшие отличия между документами разных категорий и настроить гибкие процессы классификации.
Новый интеллектуальный классификатор по внешнему виду (Image Classifier) позволяет собирать и обрабатывать визуальную информацию об изображениях документов и быстро классифицировать их. Текстовый классификатор (Text Classifier) работает с текстовой информацией на документах, в том числа анализируя смысл текста, что позволяет повысить точность классификации. Классификаторы по внешнему виду и текстовый можно использовать как отдельно, так и совместно.

Как это работает?

Классификация документов проходит в три этапа:

1. Подготовка наборов документов для обучения классификации

На этом этапе определяются классы документов. Для каждого класса подбирается несколько примеров документов для определения общих признаков.

2. Обучение классификационной модели

Информация о классах документов и соответствующих параметрах импортируется для обучения в классификационную модель (Classification Model), которая впоследствии обучается. Модель может использовать классификаторы по внешнему виду и текстовый как отдельно, так и совместно. Эффективность работы можно улучшить за счет установления баланса между полнотой и точностью данных.

3. Классификация

Все поступающие документы классифицируются согласно классификационной модели. Чтобы правильно классифицировать тип документа, определяются параметры для каждого документа, которые сравниваются с информацией, полученной на этапе обучения. Разработчики могут создавать правила, которые позволяют обновлять наборы данных для обучения и переобучать классификационную модель.

Технология позволяет получать информацию о том, к каким из заданных категорий может относиться документ и с какой вероятностью. Информацию о вероятности можно использовать для определения следующих шагов обработки, среди которых анализ и отправка документов по определенному пути.

Готовые пользовательские интерфейсы для быстрого создания приложений

ContentReader Engine содержит пять визуальных компонент на базе ActiveX, которые позволяют создать интерфейс пользователя для просмотра и предварительной обработки изображений, а также редактирования и верификации распознанного текста и отслеживания процесса. Визуальные компоненты разработаны на основании обширного опыта Контент ИИ по созданию пользовательских приложений.

Просмотр изображения (Image Viewer)

Image Viewer отображает полное изображение страницы документа и позволяет её просматривать и редактировать, а именно:

- Изменять поворот, обрезать и разделять изображения;
- Редактировать области распознавания или выбирать типы блоков — текст, таблица или штрихкод;
- Кнопки панели инструментов можно отображать или скрывать посредством кода, можно создать пользовательские кнопки.

Увеличительное стекло (Zoom Viewer)

ZoomViewer — это удобный способ рассмотреть изображение в мельчайших деталях, скорректировать расположение рассматриваемой области или сравнить неуверенно распознанные символы с увеличенным исходным изображением.

Просмотр документа (Document Viewer)

Воспользуйтесь Document Viewer, чтобы просмотреть структуру документа в целом и результаты обработки. Сохраняйте распознанный документ и открывайте в выбранном приложении. Существуют два режима просмотра:

- Детальный просмотр
- Просмотр иконок страниц

Редактор текста (Text Editor)

Text Editor позволяет подчеркивать неуверенно распознанные символы и форматировать текст. Разработчики могут контролировать размеры текстового окна, доступные кнопки и набор действий пользователя.

Проверка текста (Text Validator)

Text Validator — удобный и эффективный инструмент, который позволяет проверять неуверенно распознанные символы и правописание. Кроме того, режим позволяет проверять орфографию и просматривать проверенный текст в увеличенном масштабе. Разработчики могут полностью контролировать поведение этой библиотечной компоненты.

Визуальные компоненты поддерживают несколько языков интерфейса, также их можно локализовать на любой язык.

Функция сравнения документов

С помощью модуля сравнения документов на базе ContentReader Engine можно автоматически сравнивать две версии документа и находить в них различия, чтобы, например, проверить соответствие внесенных изменений договоренностям или убедиться в отсутствии исправлений.

Сравнение документов

Технологию сравнения документов можно легко интегрировать в системы электронного документооборота, CRM, электронные архивы и многие другие системы, чтобы расширить их возможности.

Подписание договора и других документов

В ситуациях, когда наличие ошибок в документах критично, например, при подписании договора с контрагентом, технологии позволяют быстро сравнить оригинал документа с копией и сразу обнаружить изменения.

Согласование в системах электронного документооборота

ContentReader Engine позволяет сравнить две версии одного документа. Вы сможете найти различия, обнаружить внесенные изменения или определить последнюю версию документа, даже если правки не были отражены во время редактирования и внесения изменений.

Переговорный процесс

После согласования условий сотрудничества компания отправляет бизнес-партнеру документ, например, меморандум, технический проект, а он, в свою очередь, может внести туда изменения и отправить встречный документ. Его экземпляр можно легко сравнить с оригиналом. Интеллектуальные технологии Контент ИИ помогут найти все расхождения в двух версиях документа, даже если отличия на первый взгляд не заметны.

Преимущества

Модуль сравнения поддерживает множество как текстовых, так и графических форматов, например, Word, Excel, PowerPoint, PDF, JPEG, TIFF, PNG и другие.

Технология позволяет показывать только существенные изменения: удаление, добавление, исправление текста, а также игнорировать изменения в форматировании. Все это помогает экономить время на просмотр и проверку документа.

Результаты сравнения доступны через API или могут быть экспортированы в Word (в режиме Исправлений - Track Changes) или XML.

Технологию легко интегрировать в сторонние приложения и использовать для сравнения как объемных документов, так и отдельных страниц.