Таблица: Технологии OCR для распознавания текста паспорта

Инструмент/технология	Описание и особенности
Tesseract OCR	Бесплатный движок OCR с открытым кодом, поддерживающий русский язык. Использует нейросетевую LSTM-модель для распознавания символов. Легко интегрируется (например, через Pytesseract) и работает локально. Требует предварительной обработки изображения для лучших результатов; точность на чётких печатных текстах достигает ~80–90% слов
ABBYY FineReader	Коммерческая OCR-система, известная высокой точностью на кириллическом тексте. Поддерживает сотни языков, включая русский, и использует встроенные словари для повышения точности. Предлагается в виде приложения или SDK (ABBYY FineReader Engine/FlexiCapture) для интеграции. Отличается надёжным распознаванием паспортных шаблонов, но лицензирование платное.
Google Cloud Vision OCR	Облачный OCR API от Google с поддержкой русского языка. Применяет современные модели машинного обучения; в тестах показывает наилучшие результаты по точности распознавания. Прост в использовании (REST API) и способен распознавать текст и поля. Недостатки: требуется интернет-соединение и оплата за обработку, данные передаются в облако.
Microsoft Azure OCR	Облачные сервисы Azure для OCR. Стандартный OCR поддерживает русский текст, а Form Recognizer может быть обучен на конкретном шаблоне (например, паспорт) для извлечения полей. Предоставляет структурированный вывод (ключ-значение). Требует подключения к Azure и оплачивается по использованию.
Yandex Cloud Vision	API распознавания от Яндекса, оптимизированный для русского языка. Может точно распознавать печатный текст документов. Предположительно близок по качеству к Google Vision, работает в российском облаке (важно для требований локализации данных). Также требует интернет-доступа и тарифицируется за использование.
EasyOCR	Библиотека OCR на базе глубокого обучения (PyTorch), поддерживающая >100 языков, включая русский. Включает модели детекции и распознавания текста. Легко устанавливается, работает офлайн. Качество распознавания высокое на стандартных шрифтах, но может уступать специализированным решениям (не включает спец. словари для паспортов).
Smart Engines SDK	Отечественная технология OCR на основе нейросетей, заточенная под ID-документы (паспорта, права и т.п.). Отличается высокой скоростью и точностью, а главное – полностью офлайн-работой: данные не отправляются на сторонние серверы, обработка происходит локально. Используется, например, в системах e-gate для паспортного контроля в РФ. Лицензия коммерческая.

Таблица: Методы разметки и выделения персональных данных из текста

Методы	Описание и особенности
Шаблонный подход	Поскольку формат паспорта стандартизирован, можно использовать известное расположение или метки. Например, на скане могут присутствовать печатные названия полей («Фамилия», «Имя», «Отчество», «Дата рождения», «Место рождения», «Пол»). Система может находить эти ключевые слова и извлекать следующую за ними строку как значение поля. Также возможно использовать жёстко заданные координаты областей, если скан строго выровнен по шаблону. Этот метод прост и интерпретируем, но требует, чтобы качество OCR было достаточным для чтения меток полей.
Регулярные выражения и паттерны	Для некоторых полей можно применять поиск по форме данных. Например, дату рождения можно выявить по шаблону даты (ДД.ММ.ГГГГ), серию и номер паспорта – по типичному формату из 4 цифр серии и 6 цифр номера. Такой подход дополняет шаблонный: даже если метка поля не распознана, по формату текста можно классифицировать фрагмент (например, строка из 6 цифр скорее всего номер паспорта).
Модель структурированного распознавания	Существуют специальные системы для извлечения полей из документов. Например, ABBYY FlexiCapture позволяет настроить шаблон паспорта и будет автоматически парсить скан, возвращая значения полей в структуре. Подобно этому, облачные AI-сервисы (Azure Form Recognizer, Google Document AI) могут быть обучены на образцах паспортов, чтобы модель сама научилась находить необходимые ключевые слова и значения. Эти подходы используют машинное обучение и могут быть более устойчивыми к вариациям (например, если макет немного сдвинут), но требуют набор размеченных данных для обучения.
Нейросетевое распознавание форм	Современные исследования предлагают использовать языковые и визуальные модели (например, LayoutLM и др.), которые воспринимают документ как изображение с текстом и классифицируют блоки текста по категориям полей. Для узкой задачи паспорта это, как правило, избыточно, однако упоминается как возможный метод.

Метрики оценки качества решения

1. Точность распознавания текста. Традиционно качество OCR оценивается на уровне символов или слов. Можно использовать метрики:

- *Accuracy (точность по символам)* – доля правильно распознанных символов от общего числа. Например, 98% означает, что 98 из 100 символов верны. Близкой метрикой является **Character Error Rate (CER)** – процент ошибок символов (100% - accuracy).
- *Word Accuracy* – доля полностью правильно распознанных слов. В случае паспортных данных лучше считать по полям: например, насколько часто всё ФИО распознаётся полностью правильно без ошибок.

2. Метрики информационного извлечения (по полям). Когда важна не посимвольная точность, а именно корректность извлечённых значений полей, удобно использовать метрики из области информационного поиска: **precision** (точность), **recall** (полнота) и **F1**-мера. Здесь под *найденными элементами* понимаются распознанные поля, а *релевантными элементами* – истинные поля документа. Например, если из 6 требуемых полей верно извлечены 5, а одно пропущено, recall = 5/6 ≈ 83%; если при этом лишних/неправильных полей нет, precision = 5/5 = 100%. Precision определяется как доля релевантных (правильных) данных среди всех извлечённых, а recall – как доля извлечённых релевантных данных от их истинного полного количества. F1-скор является гармоническим средним точности и полноты, давая единый показатель качества. Для оценки паспортного OCR обычно считают эти метрики по каждому типу поля и усредняют. Например, отдельно точность/полноту для распознавания дат рождения, имен, мест рождения и т.д. Такая детализация покажет, какие именно поля даются сложнее всего (например, места рождения могут распознаваться хуже из-за разнообразия топонимов).

Изучив вышеописанное, самое легкодоступное и надежное решение – это комбинация проверенного временем Tesseract и какой-то современной модели детекции (например YOLO). При должном обучении модель будет стабильно находить нужные области, даже при плохом качестве съемки.