



Optical

Character

Recognition

OCR 2.0 для банковских
документов

Интеллектуальная
система извлечения
данных

Канат Ернур
IAI



Проблема OCR в банке

- ✖ Классические OCR-системы (PyTesseract без улучшений) недостаточны для банковского сектора.
- ✖ Низкая точность на 'шумных' сканах: текст с тенями и дефектами часто теряется.
- ✖ Недостаток контекстного понимания: OCR не понимает, кто 'Seller/Продавец' и кто 'Buyer/Покупатель'.
- ✖ Негибкость: жёсткие правила для каждого типа документа → система плохо масштабируется.

Архитектура Решения

Ключевая идея: гибридная система =
OCR + LLM

1 Предобработка изображения:
очистка шума, повышение контраста,
выравнивание.

2 Базовый OCR (PyTesseract):
извлекаем весь текст.

3 Интеллектуальное извлечение
(Gemma2 LLM): анализ контекста →
структурированный JSON.

4 Постобработка и верификация:
Python + RegEx → исправление
ошибок и заполнение пропущенных
полей.

```
dpi = 300
use.preprocessing = True
OLLAMA_MODEL = "gemma2b-instruct-q4_K_M"
OLLAMA_API_URL = "http://localhost:11434/api/chat"

REQUIRED_FIELDS = [
    "document_type", "contract_number", "sign_date", "expiry_date",
    "seller", "buyer", "amount", "currency",
    "validation_status", "extraction_accuracy"
]

# Статусные сообщения
PROCESSING_MESSAGES = [
    "Подготовка ИИ-модели для анализа...",
    "Преобразование PDF в изображения...",
    "Анализ структуры документа...",
    "Распознавание текста с помощью OCR...",
    "Исправление орфографических ошибок...",
    "Извлечение ключевых полей...",
    "Валидация извлеченных данных...",
    "Формирование структурированного JSON...",
    "Подсчет метрик качества...",
    "Завершение обработки..."
]

# -----
app = Flask(__name__)
app.config['UPLOAD_FOLDER'] = UPLOAD_FOLDER
app.config['OUTPUT_FOLDER'] = OUTPUT_FOLDER

# --- OCR и обработка (ваш код) ---
def robust_preprocess(image: Image.Image) -> Image.Image:
    """Предобработка для улучшения качества перед OCR."""
    if image.mode != 'L':
        image = image.convert('L')

    if use.preprocessing:
        enhancer = ImageEnhance.Contrast(image)
        image = enhancer.enhance(2.5)
        enhancer = ImageEnhance.Sharpness(image)
        image = enhancer.enhance(3.0)

    img_array = np.array(image)

    coords = np.column_stack(np.where(img_array < 200))
    if len(coords) > 0:
        angle = cv2.minAreaRect(coords)[-1]
        if angle < -45: angle = -(90 + angle)
        else: angle = -angle
        if abs(angle) > 1:
            (h, w) = img_array.shape[1:2]
            center = (w // 2, h // 2)
            M = cv2.getRotationMatrix2D(center, angle, 1.0)
            img_array = cv2.warpAffine(img_array, M, (w, h), flags=cv2.INTER_CUBIC, borderMode=cv2.BORDER_REPLICATE)
```

Демонстрация на реальных документах

Кейс 1: Документ 6708.pdf

- Tesseract пропустил дату и сумму.
- Наша система извлекла все ключевые поля.

Кейс 2: Документ 4392.pdf

- Дата '27 февраля 2023 г.' сбивала старые системы.
- OCR 2.0 нормализовал её в формат 27.02.2023.



Метрики качества

 Как мы измеряем успех?

- Field-level Accuracy: близко к 100% (значительно выше baseline).
- Exact Match: все ключевые поля извлекаются без ошибок.
- JSON Validity: всегда возвращается валидный JSON.

OCR 2.0 для банковских документов

Загрузите PDF-документ и получите структурированные данные

Обработка документа

Результаты обработки

Общая информация		Извлеченные данные		Метрики качества	
Тип документа:	contract	Номер договора:	1	Field-level Accuracy:	86.0%
Имя файла:	2422ac2a-3e2e-4c01-b666-5552a1649973.pdf	Дата подписания:	—	Exact Match:	0.0%
Статус валидации:	partial	Дата окончания:	31.12.24	JSON Validity:	100.0%
		Продавец:	ООО «Алтай-Кабель»	Schema Consistency:	100.0%
		Покупатель:	Директора Горшкова Романа Сергеевича		
		Сумма:	182 117 993.91		
		Валюта:	RUB		

Скачать JSON

Потенциал внедрения



От MVP → к промышленному
продукту

- KYC (Know Your Customer): автоматизация проверки паспортов, анкет и договоров.
- Кредитные процессы: быстрый анализ справок о доходах и залоговых документов.
- Цифровой архив: создание структурированных баз из тысяч бумажных документов.