МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

СИСТЕМА ПЕРЕВОДА ГРАФИЧЕСКИХ ИЗОБРАЖЕНИЙ НА ОСНОВЕ ТЕХНОЛОГИЙ OCR И IMAGE INPAINTING

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Муравьева Олега Дмитриевича

Научный руководитель
доцент, к. ф.-м. н. _____ Д. В. Мельничук

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы исследования. В условиях стремительного роста объёмов мультиязычного графического контента всё более острой становится проблема его качественной локализации на русский язык. Особую сложность представляют сканированные иллюстрированные издания: текст в них расположен внутри графических выносок, а сами файлы имеют нестандартно большие размеры – до 30 000 пикселей по вертикали. Ручная локализация подобных материалов требует специальных навыков работы в графических редакторах и значительных трудозатрат.

Существующие инструменты не обеспечивают полного цикла автоматизированной обработки: библиотеки *EasyOCR* и *MangaOCR* выполняют только распознавание текста, платформы *DeepL* и *Google Translate* предоставляют API перевода без интеграции с обработкой изображений, а специализированные приложения по типу *Ballon-translator* частично автоматизируют процесс, однако не генерируют PSD-файлы и не обеспечивают полноценную обработку SFX-элементов. Итак, разработка комплексной системы, охватывающей все этапы – от загрузки изображения до формирования редактируемого PSD-файла, – является актуальной задачей с выраженной практической значимостью.

Цель работы – разработка веб-приложения для автоматизированной локализации сканированных графических изданий, реализующего полный цикл обработки: от сегментации крупноформатных изображений и OCR распознавания до формирования редактируемого многослойного PSD-файла с переведённым текстом.

Для достижения поставленной цели в работе необходимо выполнить следующие **задачи:**

— Разработать алгоритм интеллектуальной нарезки и сборки крупноформатных изображений размером до 1 800×30 000 пикселей без видимых артефактов на швах.

— Реализовать модуль OCR с поддержкой нескольких языков и сохранением пространственных координат текстовых блоков.

— Разработать классификатор текстовых элементов, разделяющий диало-

говые выноски и звукоизобразительные единицы по визуальным признакам.

— Реализовать очистку изображений: закрашивание выносок и нейросетевое восстановление фона (inpainting) для сложных областей.
— Разработать модуль автоматической вёрстки с размещением переведённого текста по координатам выносок и подбором размера шрифта.
— Сформировать многослойный PSD-файл для последующего редактирования перевода.

Практическая значимость работы заключается в разработке программного продукта, существенно сокращающего трудозатраты на локализацию иллюстрированных изданий и предоставляющего профессиональным локализаторам удобный инструмент финальной редактуры в виде структурированного PSD-файла, совместимого с Adobe Photoshop.

Объект – процесс автоматизированной локализации сканированных иллюстрированных графических изданий.

Предмет – методы оптического распознавания текста на основе мультимодальных моделей зрения и языка, алгоритмы генеративного восстановления изображений (inpainting) и подходы к автоматической типографической вёрстке применительно к задаче замены текста на изображениях.

1 Основное содержание работы. Первый раздел.

Теоретическая часть работы посвящена анализу целевой аудитории разрабатываемой системы, обоснованию архитектурных решений с использованием методологии Jobs to be Done, а также теоретическому обзору нейросетевых моделей, лежащих в основе всех модулей конвейера.

Анализ целевой аудитории проведён с использованием классической матрицы «Потребность – компетенция». Выделены три ключевых сегмента: профессиональные локализаторы и переводчики, для которых критичен многослойный PSD-файл с редактируемыми элементами; издательства и студии локализации, заинтересованные в конвейерной обработке десятков и сотен страниц в сутки; любительские сообщества переводчиков, обладающие глубоким знанием исходного языка, но не имеющие профессиональной подготовки в области графического дизайна. Максимальный приоритет назначен третьему сегменту, поскольку он одновременно испытывает острую потребность в автоматизации и не обладает ресурсами для ручной обработки. Изобразим матрицу «Потребность – компетенция» на рисунке 1.

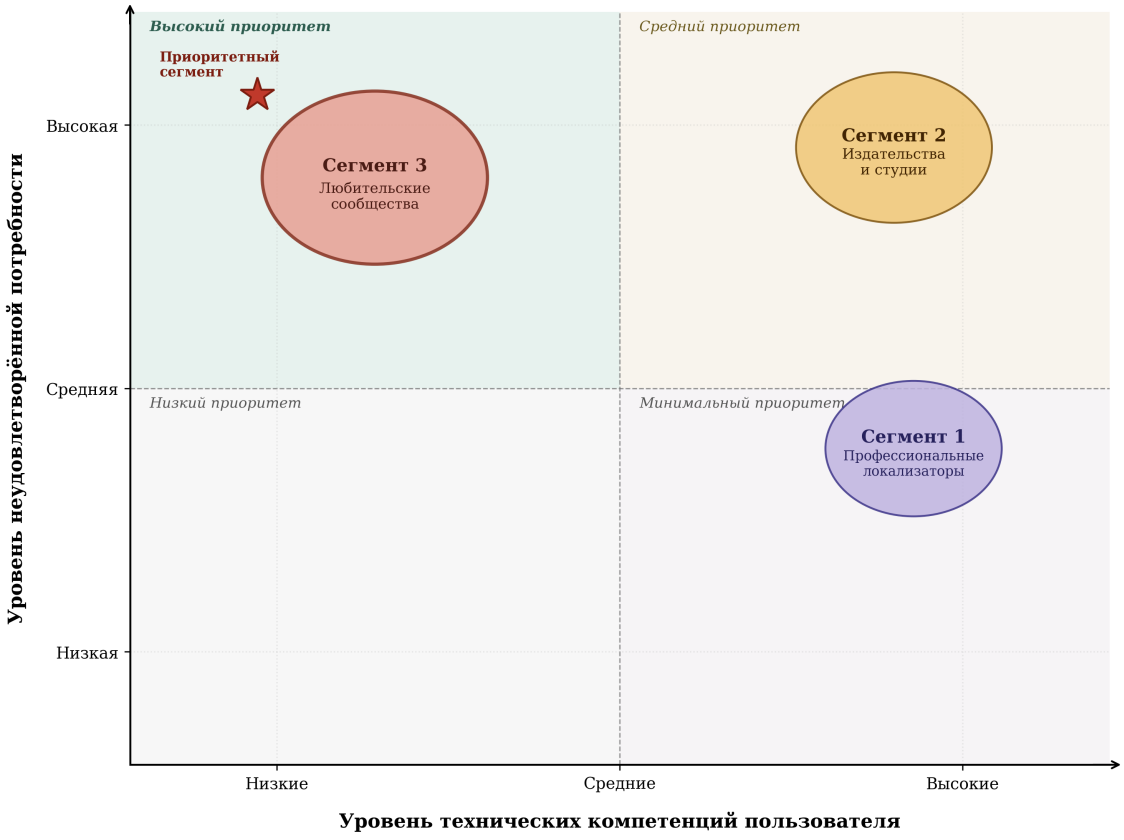


Рисунок 1 – Матрица «Потребность – компетенция»

Архитектурным фундаментом всех нейросетевых моделей, использованных в работе, является трансформер, предложенный А. Васвани и соавторами в 2017 году. Центральным элементом трансформера выступает механизм масштабированного скалярного внимания, формула которого имеет вид:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{1}$$

где Q, K, V – матрицы запросов, ключей и значений, d_k – размерность векторов ключей. Масштабирующий множитель $\sqrt{d_k}$ необходим для предотвращения экспоненциального роста значений скалярных произведений и стабилизации обучения. Формула 1 является основополагающей для всех нейросетевых моделей, задействованных в системе: механизм внимания лежит в основе мультимодального распознавания текста (Qwen-VL), диффузионной генерации (FLUX) и нейронного машинного перевода.

В разрабатываемой системе используется несколько типов нейронных сетей. Для предварительной детекции текста применяется свёрточная модель CRAFT (Character Region Awareness For Text detection), работающая на уровне отдельных символов и одинаково эффективная для корейских, японских, китайских и латинских глифов. Распознавание текста реализовано на базе мультимодальной модели Qwen3-VL – модели семейства Vision-Language Models, объединяющей визуальный кодировщик ViT, адаптер модальностей и большую языковую модель в единой архитектуре. Преимущество VLM для задачи OCR состоит в способности учитывать семантический контекст: модель не просто идентифицирует отдельные символы, но и корректирует их с учётом смысла предложения. Параллельно используется специализированная модель построчного OCR Datalab, обеспечивающая субпиксельную геометрическую точность для последующего этапа очистки.

Для генеративного восстановления фона после удаления текста (inpainting) применяется диффузионная модель FLUX. Диффузионные модели базируются на двух процессах: прямом, постепенно добавляющем гауссовский шум к изображению, и обратном, восстанавливающем изображение из шума. Обучение сводится к задаче предсказания шума, добавленного на каждом

шаге, с минимизацией функции потерь:

$$\mathcal{L} = \mathbb{E}\Big[\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2\Big] \tag{2}$$

где ε – гауссовский шум, x_t – зашумлённая версия изображения на шаге t, ε_θ – нейросетевая аппроксимация шума. Модель FLUX расширяет стандартную диффузионную модель механизмом кондиционирования на текстовый промпт и входное изображение с маской, что позволяет направлять генерацию в соответствии с контекстом.

Задача автоматической вёрстки переведённого текста внутри ограничивающего прямоугольника формализована как задача оптимизации:

$$f^* = \max\big\{\, f \in F : \mathrm{TotalHeight}\big(\mathcal{L}(T, f, w),\ f\big) \le h \,\big\} \tag{3}$$

где f^* – искомый оптимальный размер шрифта, F – множество допустимых размеров, $\mathcal{L}(T, f, w)$ – функция переноса слов с разбиением текста T на строки шириной w, TotalHeight – суммарная высота строк с учётом интерлиньяжа, h – высота ограничивающего прямоугольника. Поиск оптимального размера реализуется методом бинарного поиска, что обеспечивает логарифмическую сложность.

В качестве выходного формата выбран PSD (Photoshop Document) – индустриальный стандарт многослойных растровых изображений. Структура файла включает три категории слоёв: оригинальное изображение, очищенный фон и текстовые слои с переведёнными репликами. Такая структура обеспечивает неразрушающее редактирование: локализатор может независимо корректировать перевод, положение текста и параметры очистки. Таким образом, первый раздел формирует теоретическую базу, необходимую для дальнейшей практической реализации системы.

2 Второй раздел.

Здесь представлен процесс проектирования и реализации программного решения. По итогам разработки система приобрела вид конвейера из шести последовательно выполняемых модулей: модуля нарезки длинного холста, модуля сборки исходного холста, модуля оптического распознавания текста, модуля перевода через внешний сервис, модуля очистки изображения от исходного текста и модуля типографического размещения перевода. Координация конвейера реализована отдельно. Общая схема системы представлена на рисунке 2.

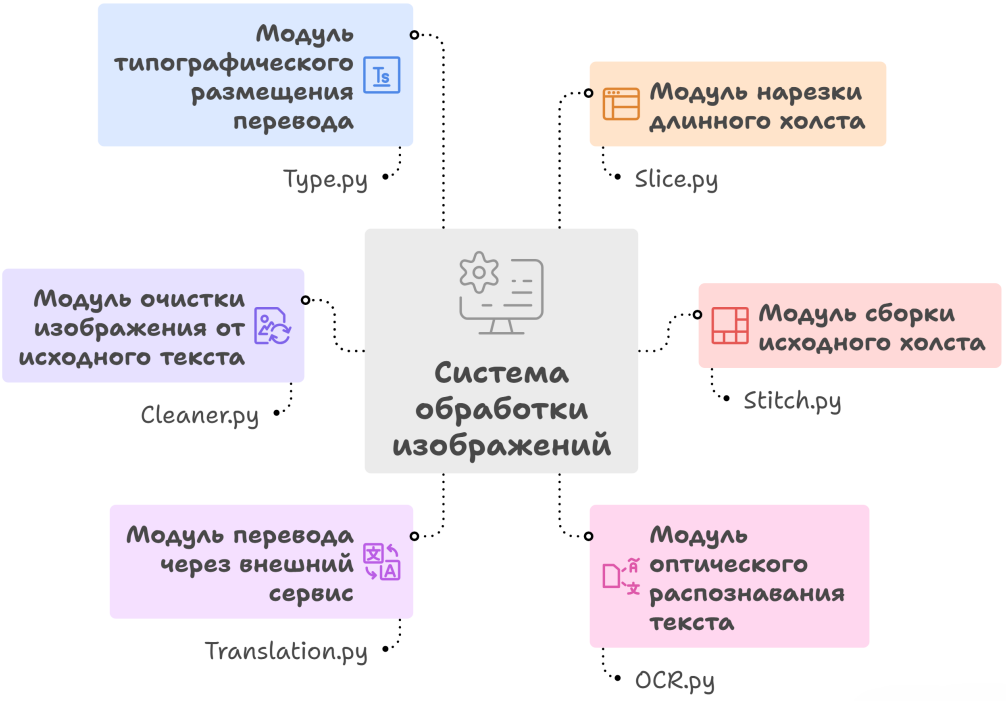


Рисунок 2 – Конвейер модулей реализованной системы

Модуль нарезки прошёл через несколько итераций. Первая версия использовала простую эвристику по средней яркости строк: точкой разреза служила середина непрерывной полосы «фоновых» строк с яркостью выше 240 из 255. На практике пороговая эвристика давала ложные срабатывания на стилизованных страницах с цветными фонами и, что критичнее, разрезала холст по строчкам текста звуковых эффектов между панелями. Итоговая версия совмещает три стратегии в каскаде: сначала ищется «линия одинаковых пикселей» (строка, в которой все пиксели идентичны), затем светлый промежуток, и только при неудаче обеих эвристик применяется жёсткое ограни-

чение по высоте с локальной проверкой моделью CRAFT в формате ONNX. Если CRAFT обнаруживает текст в верхней полосе кандидата слайса, точка разреза итеративно сдвигается выше до тех пор, пока полоса не окажется свободной от текста. На тестовом наборе из нескольких десятков страниц итоговая версия не допустила ни одного разреза через строчку текста. Результат работы кода с использованием этой стратегии представлен на рисунке 3.

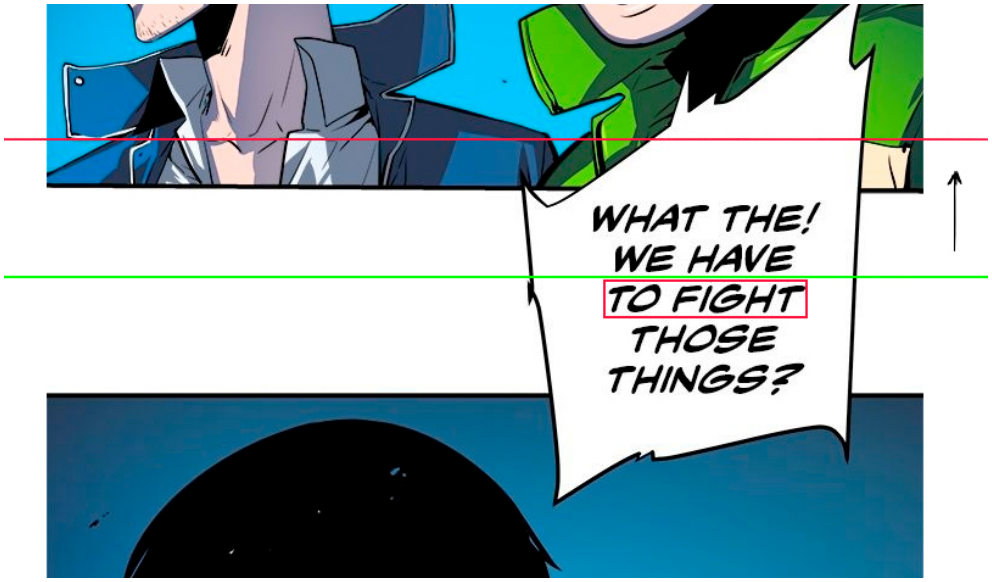
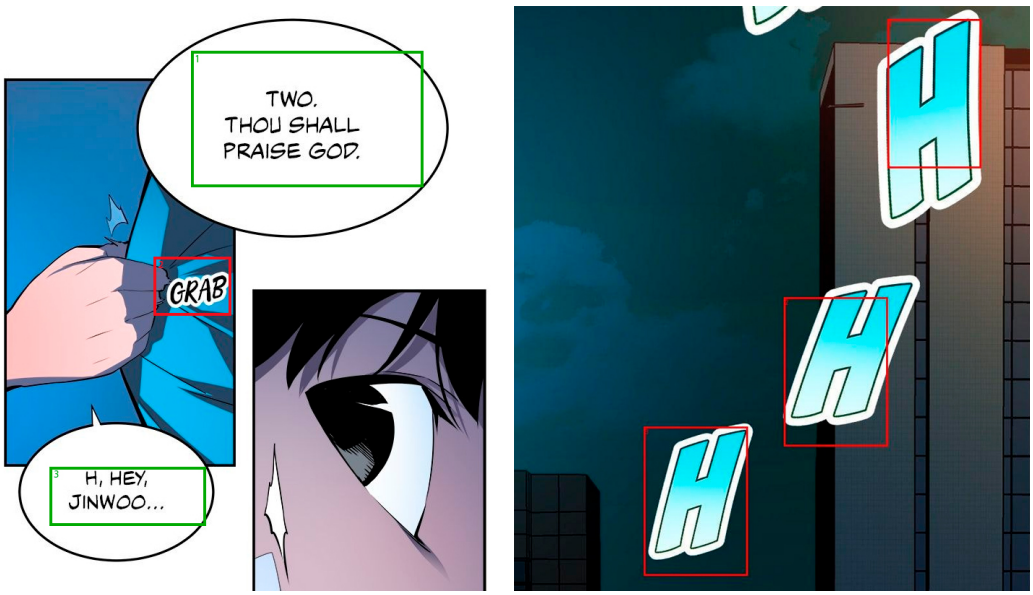


Рисунок 3 – Решение для разреза между строк текста

Модуль OCR прошёл наиболее радикальную эволюцию. Исходное решение на библиотеке EasyOCR было полностью пересмотрено по результатам трёх независимых проверок: качество распознавания стилизованных шрифтов составило 40–50% для речевых облаков и менее 10% для звуковых эффектов. Модель некорректно объединяла многострочные реплики, возвращая отдельный объект на каждую строку. В ответе отсутствовала явная информация о типе фона, необходимая для последующей очистки. Новая архитектура построена на гибридной схеме мультимодальных моделей: основное распознавание выполняет Qwen3-VL-8B-Instruct через сервис Together AI, дополнительная геометрическая разметка строк – модель Datalab. Промпт Qwen-модели жёстко специфицирует формат ответа: каждый текстовый блок описывается полями text, box_2d (нормализованные координаты 0–1000), on_white_background и type(speech / sfx / thought). Пример работы Qwen-модели представлен на рисунке 4.



(a) (b)

Рисунок 4 – Работа через Qwen3-VL-8B-Instruct

Координаты, возвращаемые Qwen, оказались семантически корректными, но геометрически приблизительными – рамка часто захватывает весь бабл вместе с белым полем. Для точной заливки на этапе очистки этого недостаточно, поэтому в систему был включён сервис Datalab OCR, специализирующийся на построчной детекции с плотными рамками. Итоговая функция process_slice_ocr сначала обрабатывает слайс через Qwen, а при наличии блоков с признаком on_white_background = true дополнительно запускает Datalab. Функция merge_ocr_outputs_for_page производит слияние результатов: для каждого «белого» Qwen-блока выбираются Datalab-линии, попадающие внутрь его bbox. Экономическая целесообразность облачных API подтверждена фактическими расходами: при объёме порядка 500 обработанных страниц затраты на Together AI составили менее 10 долларов, что многократно дешевле аренды GPU-сервера.

Модуль очистки реализует гибридную стратегию выбора метода в зависимости от контекста блока. Для блоков с признаком on_white_background = true применяется тривиальная заливка белым с использованием плотных построчных рамок Datalab, что гарантирует абсолютно чистый результат без затрагивания контура бабла. Для блоков на сложном графическом фоне (преимущественно звуковых эффектов) задействуется диффузионная модель

FLUX Kontext через сервис Replicate с промптом «Remove all visible text/SFX completely and reconstruct the underlying background. Keep original style, colors, lighting and lineart. Do not add new text, letters, symbols or watermarks.». Пример работы FLUX-инпейнтинга представлен на рисунке 5.

(a) (b)

Рисунок 5 – Работа FLUX Kontext

Преимущества гибридной стратегии перед универсальным применением FLUX состоят в трёх аспектах. Во-первых, стоимость: речевые облака составляют около 80% текстовых блоков на типичной странице, и их обработка через бесплатную локальную заливку вместо платного API сокращает суммарные расходы во много раз. Во-вторых, качество: заливка белым на белом фоне имеет нулевую вероятность артефактов, тогда как FLUX обладает некоторым процентом неудачных случаев. В-третьих, предсказуемость: две простые стратегии с чёткими критериями выбора проще отлаживать и сопровождать, чем одну сложную «универсальную» систему.

Модуль типографики решает три подзадачи: восстановление глобальных координат текстовых блоков относительно склеенного холста, разбиение переведённой строки на типографские строки заданной ширины и подбор максимального размера шрифта, при котором текст полностью помещается в ограничивающую рамку. Подбор размера реализуется бинарным поиском, что соответствует формуле 3 из теоретической части. Финальная сборка многослойного PSD-файла осуществляется через автоматизацию Adobe Photoshop посредством COM-интерфейса (библиотека comtypes): создаётся

новый документ, последовательно добавляются слой оригинального изображения, слой очищенного фона и текстовые слои с переведёнными репликами. Пример итогового результата представлен на рисунке 6.

Рисунок 6 – Результат вставки текста

Таким образом, во втором разделе последовательно реализованы все модули конвейера системы.

ЗАКЛЮЧЕНИЕ

Главная цель работы достигнута, а задачи, поставленные в начале работы, были выполнены. Основные результаты:

- Разработан алгоритм нарезки и сборки крупноформатных изображений.
- Реализован гибридный модуль OCR на базе Qwen3-VL и Datalab.
- Реализована гибридная стратегия очистки изображений с применением FLUX Kontext.
- Разработан модуль типографической вёрстки с генерацией многослойного PSD-файла.

В ходе работы было создано веб-приложение для автоматизированной локализации сканированных графических изданий, реализующее полный цикл обработки от загрузки исходного изображения до формирования редактируемого PSD-файла с переведённым текстом.