РАСПОЗНАВАНИЕ НЕСТРУКТУРИРОВАННОГО ТЕКСТА

Магистрант кафедры интеллектуальных систем Шуманов В. Е.

Научный руководитель – кандидат физ.-мат. наук, доцент Козлова Е. И.

Белорусский государственный университет

Минск, Беларусь

В Беларуси насчитывается более 25 тысяч людей с нарушениями зрения и более 28 тысяч иностранных студентов. Задача состояла в том, чтобы сделать для этих людей чтение документов, газет, веб-страниц и т.д. доступным на любом языке. Последние достижения в области компьютерных наук и обработки изображений сделали это достижимым.

Optical character recognition (OCR) – это преобразование изображений текста в цифровой формат. В настоящее время OCR латинских шрифтов имеет настолько высокую точность, что считается решенной проблемой, и многие программные пакеты предоставляют бесплатную технологию OCR для многих языков.

Рисунок 1. Визуализация алгоритма с использованием OCR.

Технология оптического распознавания символов (OCR) значительно усовершенствовалась за последние десятилетия благодаря исследовательской работе, увеличению вычислительных мощностей и передовым методам

машинного обучения. Предлагаемая модель системы помощи людям с использованием Tesseract OCR показана на рисунке 1.

Изображения рекомендуется масштабировать не менее чем до 300 DPI. Увеличение DPI выше этого значения приводит только к увеличению размера выходного файла без улучшения его качества, в то время как DPI ниже этого значения приводит к появлению шума и нежелательному результату.

Шум – это пиксели изображения, которые сильно отличаются по цвету или интенсивности от окружающих их пикселей. Хотя основные причины шума могут быть разными, очевидно, что он затрудняет распознавание символов. Удаление шума включает в себя такие методы как расширение, эрозия и размытие.

Бинаризация – это преобразование многоцветного изображения (RGB) в черно-белое. Для преобразования изображения используется функция порогового выделения библиотеки OpenCV. Tesseract OCR может выполнять бинаризацию внутренними средствами, однако, если входное изображение неравномерно темное, это приводит к некачественному результату.

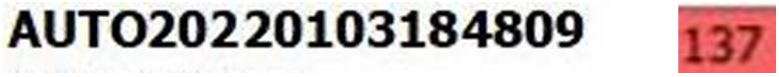

Рисунок 2. Загруженное изображение.

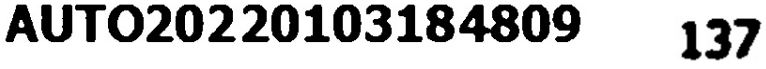

Рисунок 3. Подготовленное изображение.

Рисунок 4. Tesseract OCR.

Точность модели сильно зависит от качества входного изображения. Если говорить о точности на уровне символов, то точность модели составляет около 99% (1 из 100 символов распознается как “неопределенный”).

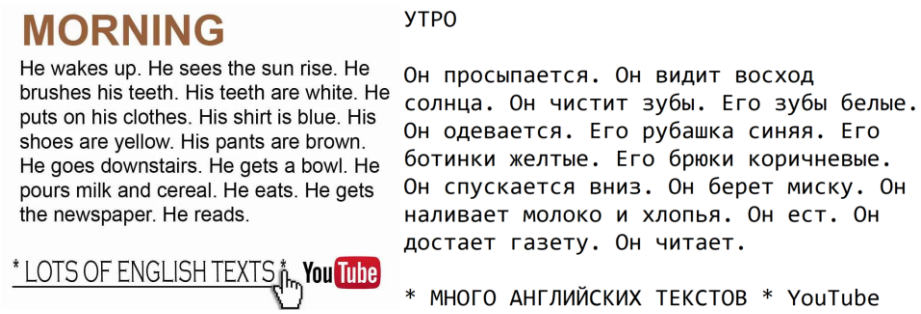

Рисунок 5. Распознанный и переведенный текст.

После того как из изображения получен подготовленный текст, как на рисунке 5, происходит преобразование текста в аудиофайл MP3 с помощью gTTS (Google Text-to-Speech).

В данной работе был реализован алгоритм для распознавания текста на основе gTTS (Google Text-to-Speech) с использованием движка Tesseract OCR. Алгоритм успешно распознает текст различных шрифтов с различных входных изображений и преобразует текст в аудиофайл, который является чрезвычайно точным. В работе использовался движок Tesseract, так как он является мощным и открытым программным обеспечением, а также не требует лицензий и инвестиций. Эксперименты проводились путем визуального сравнения тестовых примеров OCR, в итоге были получены хорошие

результаты при расчете точности модели. Точность на уровне символов составляет около 99%. Предложенная система облегчит работу с цифровыми технологиями людям со слабым зрением и с ограниченными возможностями обучения.

Литература

1. Tesseract OCR [Электронный ресурс]. – Режим доступа: https://github.com/tesseract-ocr. – Дата доступа: 17.12.2023.

2. GitHub gTTS [Электронный ресурс]. – Режим доступа: https://github.com/topics/gtts. – Дата доступа: 17.12.2023.