УДК 004.93

ГЕНЕРАТОР АУГМЕНТИРОВАННОГО НАБОРВ ИЗОБРАЖЕНИЙ ДОКУМЕНТОВ С АВТОМАТИЧЕСКИМ ПОСТРОЕНИЕМ РАЗМЕТКИ ДЛЯ ЗАДАЧ СЕГМЕНТАЦИИ И КЛАССИФИКАЦИИ

Бушуев К. Р. (Университет ИТМО),
Научный руководитель – кандидат технических наук, доцент Муромцев Д. И.
(Университет ИТМО)

В данной работе рассматривается алгоритм и архитектура программы генерации датасета документов для решения задач классификации и сегментации документов, а также оценки качества метрик алгоритмов предобработки данных. Основным преимуществом данной программы является автоматическая разметка документов по следующим признакам: тип документа (для решения задачи классификации), сегментные области документа (текст, таблицы, изображения, QR-коды, подписи и печати), параметры аугментации документа (зашумленность, угол наклона, ориентация документа). На основании результатов работы данной программы был создан набор данных для обучения гибридного алгоритма классификации документов.

Введение. В текущее время документооборот крупных компаний исчисляется миллионами цифровых копий документов, которые обрабатываются ежедневно. Одной из основных проблем является автоматическая классификация и извлечение данных из документов для передачи их в сопутствующие системы обработки данных. Для решения этой задачи используются алгоритмы машинного обучения и компьютерного зрения, однако для реализации этой задачи и оценки качества ее исполнения, требуется крупная база размеченных документов на основании которых проводится тестирования алгоритмов. В рамках данной работы предлагается способ построения датасета с произвольным количеством классов документов, а также автоматической разметкой, сегментных, классификационных и аугментационных параметров.

Основная часть. Конфигурацию стандартного документа внутреннего оборота можно представить в виде шаблона с областями с динамическими данными внутри них (ФИО, реквизиты, идентификационная информация и тд.) Для этого программу необходимо разделить на несколько модулей: Модуль автоматической генерации шаблонов документов, модуль автоматической генерации экземпляров на основании шаблонов, модуль разметки и модуль аугментации. В рамках модуля генерации шаблонов проводится вариация структуры документа с заполнением его шаблонной части и разметкой частей с вариацией контента. В модуле генерации экземпляров проводится заполнение контента документа на основании случайных данных в отведенных полях вариации контента. Модуль аугментации применяет к документу различные типы шумов (гауссовский, посторонние объекты, размытие), а также повороты и изменение ориентации документа. После работы программы создается набор из заданного числа документов с полной разметкой их структуры и аугментаций, который может использоваться на каждом шаге разработки алгоритма классификации для тестирования его качества, а так же сравнения метрик с уже существующими алгоритмами.

Выводы. В результате данной работы получена программа, способная сгенерировать набор данных для обучения модели классификации документов. Основным преимуществом полученного решения является автоматизированная разметка, однако в данный момент отсутствует возможность создания документов со структурными компонентами отличающимися от заявленных.

Бушуев К. Р. (автор) Подпись

Муромцев Д. И. (научный руководитель) Подпись