На правах рукописи

Орлов Никита Константинович

# Алгоритмы автоматической оценки качества фотографий документов

05.13.11 Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Москва – 2020

Работа прошла апробацию на кафедре Распознавания изображений и обработки текстов Федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)».

| | |
|---|---|
| Научный руководитель: | Загайнов Иван Германович, кандидат технических наук, руководитель группы компьютерного зрения в компании ABBYY. |
| Ведущая организация: | Автономная некоммерческая образовательная организация высшего образования "Сколковский институт науки и технологий" |

Защита состоится 25 декабря 2020 г., в 15-00 года на заседании диссертационного совета ФПМИ.05.13.11.001 по адресу: 141701, Московская область, г. Долгопрудный, Институтский переулок, д.9.

С диссертацией можно ознакомиться в библиотеке и на сайте Московского физико-технического института (национального исследовательского университета) https://mipt.ru/education/post-graduate/soiskateli-tekhnicheskie-nauki.php.

Работа представлена 13 октября 2020 г. в Аттестационную комиссию федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)» для рассмотрения советом по защите диссертаций на соискание ученой степени кандидата наук, доктора наук в соответствии с п. 3.1 ст. 4 Федерального закона «О науке и государственной научно-технической политике».

Общая характеристика работы

**Актуальность работы.** Научно-технический прогресс привел к тому, что большое количество методов распознавания образов и машинного зрения нашли свое практическое применение. Так уже много лет системы распознавания широко применяются в различных областях: анализ документов, медицина, банковское дело и пр. Улучшение качества интернета и развитие технологий облачных вычислений дали людям возможность пользоваться сложными системами, требующими большое количество вычислительных ресурсов с минимальными финансовыми затратами. Так, имея лишь мобильное устройство с фотокамерой, человек может оцифровывать и обрабатывать документы, получая при этом результат высокого качества прямо на мобильном устройстве, причем большая часть процесса обработки будет проходить на удаленном сервере с последующей загрузкой результата на мобильное устройство пользователя. Современные OCR (optical character recognition) - системы устроены достаточно сложно и распознавание одного документа требует различных предварительных обработок: снижения уровня шума, исправления разрешения, устранения различных смазов и дефокусировок, что характерно для снимков с камер мобильных устройств. Кроме того, необходим детальный анализ структуры документа с последующей оцифровкой его частей. Смаз, дефокусировка, блики, геометрические искажения - наиболее сложные дефекты с точки зрения их исправления, которые приводят к сильной деградации качества и, как следствие, к потере ценных данных. Кроме того, рост количества мобильных устройств увеличивает нагрузку на сервера, тем самым заставляя компании тратить большие ресурсы на масштабирование и поддержку сложной системы. Поэтому при мобильном вводе документов с целью последующего распознавания для экономии ресурсов важно иметь быстрые и эффективные методы определения качества снимка, чтобы при необходимости переснять документ до его отправки на сервер. Работа актуальна в силу необходимости получить полную предварительную информацию о распознаваемом документе и обеспечить грамотную фильтрацию пользовательских данных.

**Степень проработанности темы.** На данный момент методам построения оценки качества фотографии документов уделяется большое внимание в силу их актуальности и высокой практической применимости. Существенный вклад в развитие данной области внесли: D. Doermann, J. Kumar, H. Baird, M. Cannon, D.

Kumar, H. Lu, A. Souza, X. Peng, D. Kumar, T. Kanungo, W. Liu, Z. Wang, D. Tao и другие. Несмотря на большую научно-техническую базу, практически все работы, посвященные оценке пригодности документа к распознаванию фокусируются на анализе простых текстовых документов без каких-либо ограничений по вычислительным ресурсам. Кроме того, большинство подходов, использующих алгоритмы классического компьютерного зрения не учитывают вариативность фотографий, особенности устройств ввода и другие важные аспекты имеющие серьезное влияние на качество распознавания. Так, например, авторы некоторых публикаций строят тяжеловесные модели на основе методов глубокого обучения, скорость работы которых сравнима длительностью процесса распознавания, что делает актуальным исследование и разработку методов анализа фотографий с возможностью работы на мобильном устройстве.

**Объектом исследования** являются системы оптического распознавания

**Предметом исследования** является оценка качества пользовательских изображений для распознавания.

**Целью диссертационной работы** является разработка математических моделей и алгоритмов оценки качества документа, которые покажут высокую эффективность работы на мобильном устройстве и позволят работать с различными типами документов. Для достижения поставленной цели в диссертации ставятся и решаются следующие **задачи**:

1. Необходимо исследовать различные методы оценки качества фотографий, выявить их преимущества и недостатки.
2. Необходимо собрать выборку из наиболее популярных типов документов согласно пользовательским сценариям.
3. Необходимо разработать метод выделения и оценки качества букв на снимке, позволяющий обнаружить наиболее распространенные дефекты, которые снижают качество OCR: смаз и дефокусировка.
4. Необходимо разработать алгоритм обработки снимка документа, позволяющий строить признаковое описание фотографии документа для последующего построения модели оценки качества.
5. Необходимо построить модель машинного обучения, которая позволила бы обрабатывать документы различного типа.

6. Необходимо учесть потенциальное использование алгоритма в других задачах обработки документов и продемонстрировать его применимость.

**Методология и методы исследования.**

В диссертации используются методы компьютерного зрения, машинного обучения, глубокого обучения, математической статистики, теории алгоритмов. Для имплементации разработанных алгоритмов используется языки программирования C++ и Python, для построения моделей используются библиотеки NeoML, Keras, PyTorch.

**Научная новизна**

1. Предложен новый алгоритм автоматической оценки качества текста на фотографиях, учитывающий структурные особенности документов.
2. Разработана модель оценки качества фотографий с использованием гистограммных признаков
3. Анализ качества и скорости работы производится не только на A4 документах, но и на чеках, визитки, ID-картах.
4. Разработана новая система детектирования бликов мешающих распознаванию документа.
5. Разработана новая система классификации документов по типу, позволяющая эффективно обрабатывать документы стандартных типов и добавлять к ним новые.

**Практическая ценность и реализация результатов.** Разработанные в процессе исследований алгоритмы оценки качества, модели классификации по типу и детекторы бликов были интегрированы в программное обеспечение компании “ABBYY LLC”: Mobile Imaging SDK, FlexiCapture, FineReader.

**Основные положения, выносимые на защиту:**

1. Разработан алгоритм анализа фотографий документов с оценкой качества текста
2. Разработана модель оценки качества фотографий, которая использует гистограммные признаки.

3. Разработана новая система детектирования бликов мешающих распознаванию документа

4. Разработана новая система классификации документов по типу.

5. Для предложенных алгоритмов произведен анализ на предмет скорости и качества работы.

**Апробация работы**

Основные результаты диссертации докладывались на следующих конференциях:

1. The 8th International Workshop on Camera-Based Document Analysis and Recognition (CBDAR)
2. Workshop on Industrial Applications of Document Analysis and Recognition (WIADAR)

**Публикации**

Основные результаты диссертационной работы изложены в 2 публикациях в международном рецензируемом издании, индексируемом базой данных Scopus, в 4 патентах.

**Объем и структура работы**

Диссертация состоит из введения, 4 глав, заключения, списка литературы. Общий объем диссертации 81 страницу, включая 25 рисунков и 13 таблиц. Библиография включает 40 наименований.

**Личный вклад автора**

Автором разработаны необходимые компьютерные коды, проведены численные расчеты, обработаны данные расчетов. Все представленные в диссертации результаты получены лично автором. В случае совместных публикаций автор принимал непосредственное участие в постановке задач, планировании экспериментов и анализе результатов.

**Содержание работы**

Во **Введении** обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана

практическая значимость полученных результатов, представлены выносимые на защиту научные положения.

В **первой** главе рассматриваются различные подходы оценки качества фотографий документов, производится анализ их достоинств и недостатков, формулируются цели и задачи диссертации.

В научном сообществе задача известна как “Document Image Quality Assessment” (далее DIQA). Исторически сложилось так, что в предметной области ставились задачи связанные с наиболее популярными системами цифрового ввода данных. Например, пока доминирующей системой для оцифровки документов были сканеры, основное внимание в рамках задачи DIQA было сфокусировано на обработке ч/б документов и способах оценки их качества. Затем, с появлением альтернативных методов ввода данных, например, мобильных телефонов и с ростом качества получаемых при помощи мобильного устройства фотографий стала актуальной задача оценки качества фотографий документов. Первоначальные методы оценки основывались на градиентных методах и на оценках качества различных элементарных структурных частей документа: букв, слов и др. В результате роста популярности методов глубокого обучения, данные подходы были применены и к задаче DIQA. Стоит отметить, что возможность применения методов машинного обучения к данной задаче обусловлена доступностью двух публичных выборок, на которых можно сравнить различные подходы: SmartDoc-QA и SOC. SOC - очень важная выборка для предметной области, т.к. это первый набор данных, который позволил привнести метрическую компоненту в процесс сравнения результатов работы предлагаемых методов. SmartDoc-QA был опубликован позже и содержит существенно больше изображений и отличается высокой вариативностью условий съемки. Кроме того, исследователю доступна полная информация о процессе фотографирования (модель мобильного устройства, положение руки и др.). Основным недостатком публичных данных является низкая вариативность с точки зрения типов документов, их контента и фона. (см. Рисунок 1)

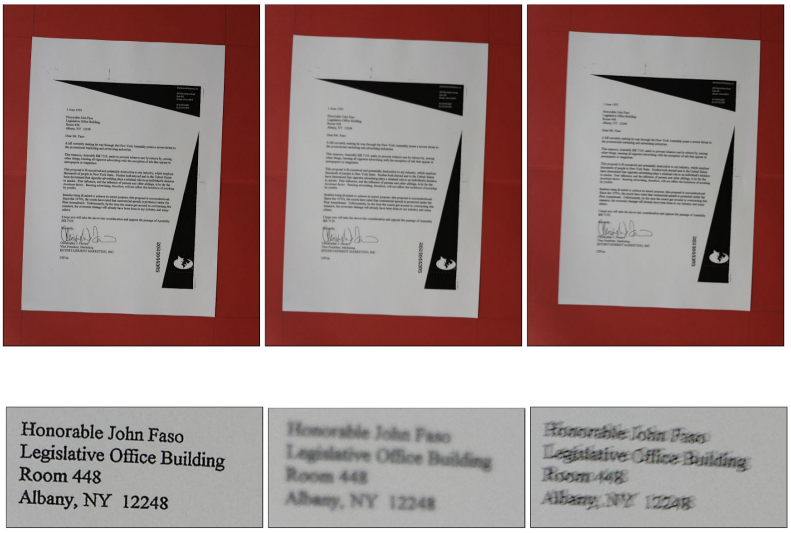


Рисунок 1.

В процессе решения задачи была сформирована база ABBYY.QA, которая пытается частично покрыть недостатки публичных баз данных, чтобы оценить качество предложенных методов на данных близких к пользовательским: 4 типа документов (A4, визитки, ID-карты, чеки) на корейском, японском, китайском и английском языках. Общее соотношение числа изображений и размеры серий приведены в таблице 1.

| | Общее число изображений | Размер серии (шт.) | Число уникальных документов |
|---|---|---|---|
| SOC | 175 | 6-8 | 25 |
| SmartDoc-QA | 4260 | 100-200 | 30 |
| ABBYY.QA | 14556 | 4-12 | 1636 |

Таблица 1.

На данный момент существует два основных направления: с использованием классических методов компьютерного зрения и нейросетевые подходы. Стоит отметить, что наилучшее качество среди классических методов может быть получено при помощи метода $\Delta DOM$, предложенного J. Kumar, который может

быть портирован на мобильное устройство. Нейросетевые подходы фокусируются на максимизации качества и целевых метрик:

$$PLCC = \frac{\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i-\overline{y})^2}} \text{ и } \quad SROCC = 1 - \frac{6\sum_{i=1}^{N}(u_i-v_i)^2}{N(N^2-1)},$$

где $x_i$, $y_i$ - предсказанная и целевая оценки качества, а $u_i$ и $v_i$ - ранги для соответствующих оценок качества. При этом важно заметить, что с течением времени растет число весов в нейросетях, что приводит к снижению скорости обработки одного документа тем самым снижая шанс на практическую применимость. Так обработка одного документа варьируется от *нескольких секунд* до *минуты*. Кроме того, в статьях используются подходы к разделению выборки на обучающую и тестовую части, которые приводят к тому, что один и тот же документ может быть в обеих частях выборки, что меняет предметную область задачи, т.к. в рамках диссертации мы рассматриваем no-reference методы, которые предполагают отсутствие априорных знаний об обрабатываемых документах.

Во **второй** главе разрабатывается основной подход к анализу фотографии документа для последующей оценки его качества: описан алгоритм обработки документа, методы оценки его структуры. Также описаны методы извлечения текста, алгоритм оценки резкости текста, методы оценки размера букв, межстрочных и меж буквенных интервалов и других свойств текста.

Общая идея метода заключается в разбиении фотографии документа на небольшие квадратные блоки фиксированного размера (32-64 пикселя). Далее, каждый блок будет рассматриваться отдельно и подлежит классификации на наличие/отсутствие текста в нем. В рамках диссертации рассматриваются два метода извлечения текста:

1. Пороговый: рассматривается динамический диапазон блока $D = I_{max} - I_{min}$ и некоторый порог $T$. Если $D > T$, то блок считается текстовым, иначе нет. Критерий кажется достаточно простым, но в рамках алгоритма этого достаточно, т.к. в дальнейшем предполагается бинаризация, которой возможна дополнительная фильтрация.
2. Нейросетевой: решается стандартная задача детектирования текста. В результате в соответствие каждому изображению ставится множество

текстовых блоков, для которых необходимо оценить качество. Однако, чтобы удовлетворить требованиям по скорости и качеству, необходимо разработать отдельную модель, которая смогла бы увеличить полноту первого подхода. Пример работы предлагаемого метода представлен на рисунке 2. Основная идея заключается в разбиении всего изображения на блоки разного размера (синие квадраты на рисунке 2). Если блоки описывают один регион текста, то они будут обязательно связаны (красные отрезки на c) и d)). При этом важно понимать, что регион текста можно покрыть как маленькими, так и большими блоками. Соединим их ссылками тоже. Таким образом нам достаточно разработать модель, которая будет для каждого масштаба блоков решать задачу классификации на предмет наличия текста и оценивать вероятность наличия связей между блоками. Время работы такой сети на мобильном устройстве составляет порядка *200 миллисекунд* на IPhone XS при максимальном размере *~4 мегабайта* без существенных потерь по качеству. Сравнимые по качеству методы работают *несколько секунд на CPU* и занимают порядка *1 гигабайта* памяти.

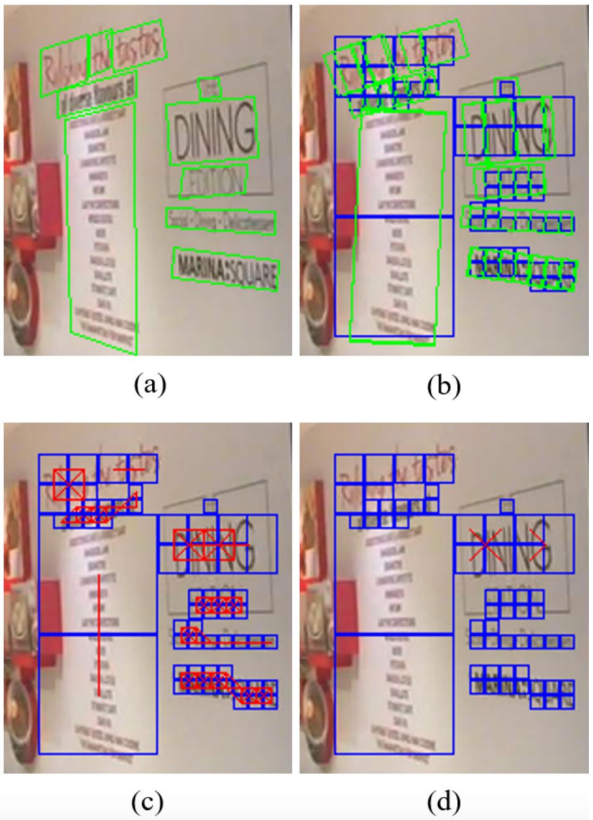


Рисунок 2.

Затем оценивается резкость пикселей на границах букв, но с применением базовых техник анализа документов. Соответственно, для этого нужно оценить число резких и контурных пикселей. Воспользуемся быстрой бинаризацией документа, чтобы оценить число контурных пикселей. Порог бинаризации выбирается, как середина динамического диапазона $D$ для рассматриваемого блока. Для построения более точной оценки можно использовать обученный ранее детектор текста, который позволит повысить полноту. Остается лишь оценить количество резких пикселей. Для большей эффективности число резких пикселей предлагается оценивать через гистограммы производных: строятся гистограммы для вертикального, горизонтального и двух диагональных направлений. Также все гистограммы суммируются в одну общую. Далее делается фундаментальное предположение для данного метода: “Почти наверное, используется не самая точная бинаризация, поэтому те пиксели, которые удалось найти точно являются резкими.” Тогда в глобальной гистограмме можно найти порог производной $t_1$, чтобы считать пиксель резким. Будем искать именно в глобальной гистограмме, т.к. известно лишь общее число контурных пикселей. Затем получаем число резких пикселей из остальных гистограмм. Заметим, что быстрая бинаризация дает лишь оценочное число контурных пикселей, которое нуждается в корректировке. На самом деле можно показать, что это значение напрямую связано с оценками толщины контура и является характерной для системы распознавания константой. Поэтому зафиксируем это как параметр алгоритма и будем искать его в процессе кросс-валидации: $t_2$ - порог производной для числа контурных пикселей, такой что $t_1 = coeff * t_2$. Таким образом имея оба порога, проинтегрируем каждую гистограмму производных и получим оценки числа резких и контурных пикселей. Будем использовать их соотношение, как дополнительные признаки в процессе построения моделей. Не менее важной частью алгоритма является оценка свойств найденного текста, т.к. на качество распознавания потенциально влияют не только резкость контуров букв, но и их размер, расстояние между ними, расстояние между строками. Все эти оценки можно построить на основе run length encoding представления бинарного изображения, которое позволяет рассмотреть бинарное изображение, как композицию белых и черных линий (штрихов) толщины в один пиксель. Для

произвольного текстового блока можно построить гистограммы белых и черных штрихов и оценить различные свойства текста по ним.

В **третьей** главе описываются основные признаки QA.Features извлекаемые при помощи предложенных ранее методов, описывается используемый алгоритм градиентного бустинга, способ его настройки, а также приводятся результаты на основных выборках описанных в первой главе. Также приводится сравнительный анализ, выявляются достоинства и недостатки предложенного алгоритма. Основные результаты представлены в таблице 2, 3, 4.

| | SROCC | PLCC |
|---|---|---|
| $\Delta DOM$ | 0.59 | 0.58 |
| Ours* (QA.Features + Gradient boosting) | 0.74 | 0.78 |

*Таблица* 2. (Результаты на ABBYY.QA)

| | SROCC | PLCC |
|---|---|---|
| $\Delta DOM$ | 0.596 | 0.575 |
| Ours* | **0.68** | **0.78** |
| Ours** | **0.7** | **0.85** |
| Camera Captured DIQA | 0.78 | 0.927 |
| Textline based framework | 0.746 | 0.719 |
| Patch-based deep learning approach | 0.931 | 0.965 |

*Таблица* 3. (Результаты на SmartDoc-QA)

| | SROCC | PLCC |
|---|---|---|
| $\Delta DOM$ | 0.95 | 0.84 |
| Ours* | **0.93** | **0.94** |
| CG-DIQA | 0.856 | 0.906 |
| Camera Captured DIQA | 0.957 | 0.936 |
| Textline based framework | 0.856 | 0.914 |

| | | |
|---|---|---|
| Patch-based deep learning approach | - | - |

*Таблица 4. (Результаты на SOC)*

В процессе экспериментирования на выборке SmartDoc-QA рассматривались два вида разбиений данных на обучающую и тестовую выборки, чтобы оценить потенциальную разницу в качестве других результатов в предметной области. (Ours** - полностью случайное разбиение на обучающую и тестовую выборки, Ours* - разбиение с контролем непопадания одного и того же документа в обе выборки или же разбиение по сериям снимков). Стоит отметить, что текущий state-of-the-art метод “Camera Captured DIQA” не учитывает тип документов при разбиении, что может быть выяснено как из текста статьи, так и из кода предоставленного авторами. Относительно других методов предложенный уступает по коэффициенту Пирсона (мы не получаем точную оценку OCR), но выдает сравнимое качество по критерию Спирмена, что позволяет использовать предложенный метод для ранжирования кадров и последующего выбора лучшего кадра без потерь качества относительно других методов. Также стоит отметить эффективную реализацию метода на C++, которая без дополнительных трудозатрат была портирована на мобильное устройство. Среднее время для вычисления признаков с последующим запуском модели градиентного бустинга составляет не более *45 миллисекунд*, а размер модели *0.3 мегабайта.*

В **четвертой** главе предлагаются альтернативные применения разработанному методу: применение к задачам классификации фотографий документов по типам и к задаче детектирования бликов.

Задача классификации изображений является одной из классических задач компьютерного зрения и машинного обучения. В последнее время наиболее популярным подходом к построению моделей классификации изображений является обучение сверточной нейронной сети. В случае, если размер выборки достаточно мал и обучение со случайной инициализацией весов затруднено, работа ведется с заранее обученной моделью на каком либо релевантном наборе данных. В итоге методом сбора дополнительных данных и аккуратным дообучением модели можно добиться неплохих результатов. Однако, с развитием мобильных технологий стала актуальной задача переноса решения на мобильное устройство. Для некоторых моделей нейросетей это достаточно трудоемкая задача. В рамках данного раздела

мы можем сравнить стандартные нейросетевые модели и классификатор на признаках QA.Features для оценки качества фотографии документа из глав 2 и 3. Оценивание будем производить на выборке ABBYY.QA, которая содержит 4 различных типа документов. В среднем предложенный подход работает с accuracy 0.88. Альтернативные методы выдают качество 0.9 для MobileNet.v2 и 0.93 для ResNet-50. Подробное качество для каждого метода приведено на рисунке 3.

QA.Features ResNet-50 MobileNet.v2

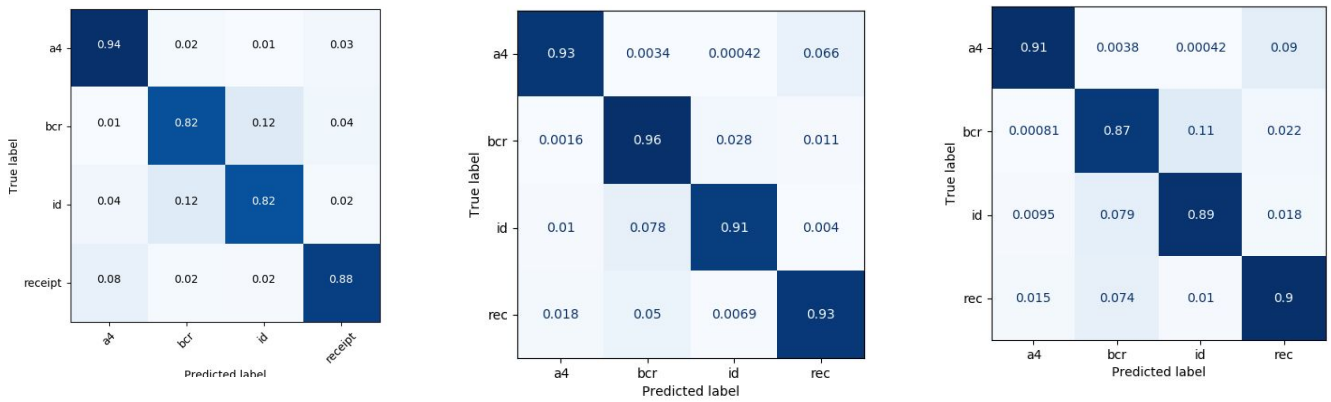


Рисунок 3.

Основная причина отставания предложенного метода по качеству - отсутствие обработки для нетекстовых регионов документа, что усложняет отличие некоторых классов таких как: визитки и id-карты. В то же самое время, предложенный подход с использованием признаков из DIQA может быть легко портирован на мобильное устройство. Время работы алгоритма составляет не более *50 миллисекунд*, а размеры модели не превышают *200 килобайт* (показатели зависят от выборки и результатов кросс-валидации). Размер сравниваемых в данном подходе нейросетей составляет *50 мегабайт* для ResNet-50 и *13 мегабайт* для мобильной сети MobileNet-v2. На первый взгляд может показаться, что размер мобильной сети достаточно мало, но важно помнить, что кроме модели также необходим фреймворк для запуска модели, что добавляет от *3* до *15 мегабайт* к финальному размеру (зависит от фреймворка). Размер кодовой базы для предложенного метода не превышает *1 мегабайта* (зависит от настроек компилятора).

Финальная часть главы посвящена применению предложенного метода в задаче детектирования бликов. Аналогично сравниваются два стандартных подхода к локализации: нейросеть для задачи сегментации обученная с использованием

бинарной кросс-энтропии и нейросеть на признаках вычисляемых в процессе вычисления QA.Features. Для повышения точности детектора предлагается не вычислять статистики по гистограммам, а использовать сами гистограммы, сформировав тензор и подав его на вход нейросети, чтобы в процессе обучения модели можно было вычислить требуемые признаки. Результаты сравнения методов приведены в таблице 5. В для оценки качества использовались стандартные метрики для бинарной классификации: точность, полнота и f-мера.

| Название метода | F-Measure | Precision | Recall |
|---|---|---|---|
| Сеть на признаках | 0.740 | 0.744 | 0.736 |
| U-net | 0.798 | 0.767 | 0.841 |

*Таблица 5.*

В результате экспериментов было выявлено, что рассматриваемый метод проигрывает классическому подходу при одинаковом размере моделей (~*200 килобайт* без учета фреймворка для запуска), но является более быстрым (более *10-15%*) и в большинстве случаев детектирует те же блики, но с некоторыми ошибками локализации краев.

В **Заключении** сформулированы основные результаты диссертации.

**Основные результаты**

В рамках диссертационной работы были решены следующие задачи:

1. Был разработан метод выделения и оценки качества букв на снимке, позволяющий обнаружить наиболее распространенные дефекты, которые мешают OCR: смаз и расфокусировка. Также были реализованы дополнительные инструменты для оценки свойств текста, оценки общей структуры документа.

2. Была собрана выборка изображений ABBYY.QA из наиболее популярных типов документов согласно пользовательским сценариям, для всех снимков было подсчитано качество распознавания и приведены другие статистики, собранные по выборке.

3. Был разработан общий алгоритм обработки снимка документа, позволяющий строить признаковое описание документа для последующего построения модели оценки качества.

4. Была построена и портирована на мобильное устройство модель машинного обучения, которая позволила обрабатывать документы различного типа. Предложенный подход был сравнен с другими алгоритмами в предметной области и показал высокую эффективность.

5. Разработанный алгоритм был применен к другим задачам OCR: классификация документов по типам и детектирование бликов, что позволило продемонстрировать вариативность предложенного подхода.

6. Предложенный метод работает достаточно быстро, чтобы запускаться в видео потоке и показывать оценку качества текущего кадра пользователю в реальном времени, что позволяет теоретически сократить объем пересылаемых данных на 20% и тем самым разгрузить сервера для облачных вычислений.

**Патенты и публикации автора по теме диссертации**

1. Method and system that determine the suitability of a document image for optical character recognition and other image processing / I. Zagaynov, V. Loginov, N. Orlov, 2017-01-17, RU2608239

2. Automatic definition of set of categories for document classification / N. Orlov, K. Anisimovich, 2019-10-19, RU2701995

3. D. Rodin and N. Orlov, Fast Glare Detection in Document Images; 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), Sydney, Australia, 2019, pp. 6-9, doi:10.1109/ICDARW.2019.60123.

4. FaSTExt: Fast and Small Text Extractor / A Filonenko, K Gudkov, A Lebedev, I Zagaynov, N Orlov, 2019 International Conference on Document Analysis and Recognition Workshops, IEEE, vol 4. pp. 49-54

5. Method and system that determine the suitability of a document image for optical character recognition and other image processing / I. Zagaynov, V. Loginov, N. Orlov, 2017-10-12, US15165512

6. Automatic definition of set of categories for document classification / N. Orlov, K. Anisimovich, 2019/9/26, US15939092

Орлов Никита Константинович
Алгоритмы автоматической оценки качества фотографий документов

Автореф. дис. на соискание ученой степени канд. техн. наук
Подписано в печать __ .___ .__ Заказ №_____
Формат 60×90/16. Усл. печ. л. 1.5. Тираж 20 экз
Типография _____