ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР «ИНФОРМАТИКА И УПРАВЛЕНИЕ» РОССИЙСКОЙ АКАДЕМИИ НАУК, ИНСТИТУТ СИСТЕМНОГО АНАЛИЗА

На правах рукописи
УДК 303.732

Булатов Константин Булатович

Методы, модели и алгоритмы комбинирования и останова в системах распознавания в видеопотоке

Специальность 05.13.01 —
«Системный анализ, управление и обработка информации
(информационно-вычислительное обеспечение)»

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:
канд. тех. наук
Арлазаров Владимир Викторович

Москва — 2019

Оглавление

Стр.

Стр.

Введение

Системы анализа и распознавания документов занимают значительное место в таких областях науки, как искусственный интеллект, теория принятие решений, и распознавание образов. Большой вклад в развитие данного научного направления внесли отечественные и зарубежные ученые М.А. Айзерман, В.Л. Арлазаров, Э.М. Браверман, Ю.В. Визильтер, И.Б. Гуревич, С.Ю. Желтов, Ю.И. Журавлев, А.Б. Мерков, А.Б. Петровский, В.А. Сойфер, Ян Лекун (Франция), Чэн-Линь Лю (КНР), Коити Кисэ (Япония), Джеффри Хинтон (Канада) и другие.

Использование смартфонов и планшетных компьютеров для решения задач оптимизации бизнес-процессов в корпоративных системах и процессов в системах государственного управления привели к новому витку развития систем компьютерного зрения, оперирующих на мобильных устройствах. Повышенный интерес к реализации корпоративного делопроизводства на основе мобильного документооборота, а также необходимость осуществления ввода документов в условиях с неконтролируемыми условиями съемки, повышают требования к системам распознавания, автоматического ввода и анализа документов с использованием мобильных устройств.

Изображения, полученные при помощи мобильных устройств, обладают рядом характерных особенностей и искажений, таких, как недостаточное разрешение, недостаточная либо неравномерная освещенность, смазывание, дефокусировка, блики на отражающей поверхности плоских объектов и другими. Подобные особенности входных изображений повышают требования к мобильным системам оптического распознавания и создают потребность в новых методах и алгоритмах, обладающих большей устойчивостью. Разработке методов распознавания образов, учитывающих особенности малоформатных цифровых камер, посвящены работы таких авторов, как Д.П. Николаев, О.А. Славин, Д.С. Ватолин, V. Lepetit, T. Geraud, R. Manmatha, D. Doermann, X. Bai, D. Karatzas, M. Iwamura и других. В то же время недостаточно изученными являются модели и методы использования видеопотока в качестве цифрового представления распознаваемого объекта, и методы повышения качества систем оптического распознавания путем использования множества гомогенных наблюдений распознаваемого объекта. Таким образом, дальнейшее исследова-

ние и развитие математических моделей и методов использования видеопотока в качестве цифрового представления объекта в контексте систем оптического распознавания является актуальным.

Основные результаты диссертации были получены в процессе выполнения работ по следующим научным грантам РФФИ:

– № 18-07-01387 – «Модели и методы построения систем оптического распознавания видеопотока с использованием обратных связей, функционирующим в в условиях ограниченных вычислительных ресурсов»;

– № 17-29-03370 – «Методы биометрической идентификации в реальном времени на мобильном устройстве по удостоверяющей фотографии»;

– № 17-29-03170 – «Исследование быстродействующих методов и алгоритмов обработки изображений и оптического распознавания для использования в мобильных устройствах с ограниченной вычислительной производительностью»;

– № 15-07-06520 – «Методы контроля подлинности документов и их фрагментов в гибридных системах обработки, передачи и хранения документов»;

– № 14-07-00730 – «Математическое моделирование шумовых помех при распознавании»;

– № 13-07-12172 – «Распознавание документов удостоверяющих личность с помощью веб камер и камер мобильных устройств».

Целью данной работы является разработка математических моделей, методов улучшения характеристик систем распознавания объектов в видеопотоке путем комбинирования результатов обработки множества входных наблюдений.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Провести анализ принципов построения современных систем распознавания документов;

2. Построить математическую модель системы распознавания объекта в видеопотоке, позволяющую исследовать качественные характеристики результата и время, необходимое для его получения;

3. Исследовать влияние характеристик входных данных на выбор оптимальной стратегии комбинирования результатов распознавания одиночных изображений;

4. Разработать алгоритм комбинирования результатов оптического распознавания строкового объекта и провести экспериментальный анализ его характеристик;

5. Разработать метод останова процесса распознавания объекта в видеопотоке в рамках построенной математической модели системы;

6. Разработать алгоритм останова процесса распознавания строкового объекта и провести экспериментальны анализ его характеристик;

7. Реализовать разработанные методы и алгоритмы для их внедрения в промышленные системы распознавания объектов в видеопотоке.

Методология и методы исследования основаны на системном анализе, математическом моделировании, математической статистике и теории принятия решений.

Основные положения, выносимые на защиту:

1. Построена математическая модель системы распознавания объекта в видеопотоке с блоком комбинирования результатов распознавания одиночных кадров и с блоком принятия решения об останове;

2. Экспериментально показано преимущество правила максимальной оценки как стратегии комбинировании покадровых результатов классификации объекта в видеопоследовательностях, не содержащих ошибок локализации и сегментации объекта;

3. Разработан алгоритм комбинирования результатов распознавания строкового объекта, учитывающий альтернативные варианты классификации отдельных символов;

4. Разработан метод останова процесса распознавания объекта в видеопотоке на основе порогового отсечения оценки ожидаемого расстояния между текущим и следующим интегрированными результатами распознавания;

5. Разработан алгоритм моделирования интегрированного результата распознавания на следующем шаге и вычисления оценки расстояния между текущим и следующим интегрированными результатами для применения метода останова.

Научная новизна:

1. Предложена новая математическая модель системы распознавания объекта в видеопотоке, позволяющая проводить совместное исследование качественных характеристик результата распознавания и времени, необходимого для получения результата;

2. Выполнено оригинальное исследование влияния характеристик входных данных на выбор оптимальной стратегии комбинирования покадровых результатов, применительно к задаче классификации объекта в видеопоследовательности;

3. Разработан новый алгоритм комбинирования результатов распознавания строкового объекта, учитывающий альтернативные варианты классификации отдельных символов;

4. Предложен новый метод останова процесса распознавания произвольного объекта в видеопотоке, рассматривающий данный процесс как монотонную задачу останова и основывающийся на оценке ожидаемого расстояния между текущим и следующим интегрированными результатами;

5. Разработан новый алгоритм останова процесса распознавания строкового объекта в видеопотоке, основанный на оценке ожидаемого расстояния между текущим и следующим интегрированными результатами, вычисляемой по накопленным наблюдениям.

Практическая значимость. Разработанная в рамках диссертации модель системы распознавания объектов в видеопотоке, а также разработанные методы и алгоритмы комбинирования результатов распознавания строковых объектов и останова процесса распознавания были реализованы в виде программных компонентов и внедрены в программное обеспечение «Smart 3D OCR MRZ» и «Smart PassportReader» компании ООО «Смарт Энджинс РУС», а также «Smart IDReader» компании ООО «Смарт Энджинс Сервис». Данные продукты интегрированы в информационную инфраструктуру ряда коммерческих организаций, а также в ряд информационных решений государственных структур Российской Федерации.

Достоверность полученных результатов подтверждается согласованностью разработанных алгоритмов, методов и математических моделей с экспериментальными результатами, представленными в работе, успешной апробацией результатов и внедрением в коммерческие системы распознавания документов.

Апробация работы. Основные результаты работы докладывались на следующих семинарах и конференциях:

1. 7th International Workshop on Camera Based Document Analysis and Recognition (CBDAR 2017), Киото, Япония, 2017;

2. 10th International Conference on Machine Vision (ICMV 2017), Вена, Австрия, 2017;

3. Международный научно-исследовательский семинар «Анализ и понимание изображений (Математические, когнитивные и прикладные проблемы анализа изображений и сигналов)», Москва, Россия, 2019.

Личный вклад. Все результаты, изложенные в диссертации, принадлежат лично автору. В совместных работах автор принимал непосредственное участие в выборе направления и задач исследований, в построении математических моделей и обсуждении результатов экспериментальных исследований.

Публикации. Основные результаты по теме диссертации изложены в 14 публикациях, в том числе: 6 изданы в журналах, рекомендованных ВАК, 3 – в сборниках трудов конференций (входящих в международные базы цитирования Scopus и Web of Science), 2 патента на полезную модель и 3 свидетельства о государственной регистрации программы для ЭВМ.

Объем и структура работы. Диссертация состоит из введения, четырех глав и заключения. Полный объем диссертации составляет 109 страниц, включая 18 рисунков и 7 таблиц. Список литературы содержит 139 наименований.

Краткое содержание глав. Первая глава посвящена анализу принципов построения современных систем распознавания документов. Рассматривается автоматический ввод документов как одна из основных задач, возникающих в рамках электронного и мобильного документооборота. Описаны основные компоненты таких систем и их свойства. Показано, что современные работы, связанные с автоматическим вводом и распознаванием документов на мобильных устройствах, рассматривают фотографию документа как его электронное представление и отмечают трудности, связанные с подготовкой образа документа к распознаванию и с самим распознаванием.

Во второй главе предложена новая математическая модель системы оптического распознавания объекта в видеопотоке с блоком комбинирования результатов распознавания объекта на одиночных изображениях и с блоком останова процесса распознавания. Предложена постановка задачи распознавания в рамках такой системы.

Третья глава посвящена разработке алгоритма комбинирования (интеграции) результатов распознавания строкового объекта в видеопотоке в рамках модели результата, учитывающей альтернативные варианты классификации отдельных символов. Описана постановка задачи, формальное описание алго-

ритма, а также представлены результаты сравнительного экспериментального исследования предложенного алгоритма и алгоритма ROVER.

В четвертой главе предложен новый метод останова процесса распознавания объекта в видеопотоке на основе порогового отсечения оценки ожидаемого расстояния между текущим и следующим интегрированными результатами, и представлен новый алгоритм останова распознавания строчного объекта. Рассмотрена формальная постановка задачи останова процесса распознавания, предложен метод, полученный путем рассмотрения задачи как монотонной задачи останова. Представлены результаты сравнительного экспериментального исследования предложенного алгоритма и других правил останова, предложенных ранее для подобных задач. Показано преимущество предложенного алгоритма перед другими методами.

Глава 1. Анализ принципов современных систем распознавания документов

1.1 Автоматический ввод документов

Документационное обеспечение управления производством, или делопроизводство [1] является неотъемлемой частью любого предприятия и заключается в создании, учете, хранении и организации движения документов. Комплекс работ по организации движения документов в организации называется документооборотом, в него входит ввод, прием, регистрация, рассылка, контроль исполнения, формирование дел, хранение и повторное использование и т.п. Для автоматизации делопроизводства на предприятиях вводится электронный документооборот – единый механизм по работе с документами в электронном виде.

Согласно официальной формулировке в законодательстве РФ документом называется материальный носитель с зафиксированной на нем в любой форме информацией в виде текста, звукозаписи, изображения и (или) их сочетания, который имеет реквизиты, позволяющие его идентифицировать, и предназначенный для передачи во времени и в пространстве в целях общественного использования и хранения [2]. Близким к понятию документа, особенно в контексте электронного документооборота, является понятие формы как набора информационных полей (реквизитов), имеющего определенную логическую структуру, а также логическое и визуальное представление.

Одним из аспектов электронного документооборота является автоматический ввод документов – метод автоматизированного ввода данных с использованием заранее определенных шаблонов и конфигураций документов. Автоматический ввод документов возник как альтернатива ручному вводу для минимизации типографических ошибок и временных затрат. Типичный технологический процесс автоматического массового ввода документов на предприятии можно описать следующими этапами:

1. Распределение потока документов на пакеты для отдельной обработки.
2. Оцифровка документов в обрабатываемом пакете, т.е. преобразование документа с бумажного или иного физического носителя в электронный вид. В случае документов на бумажных носителях данный этап

чаще всего представляет собой сканирование пакета документов при помощи высокоскоростного промышленного сканера.

3. Подготовка оцифрованных документов к распознаванию, т.е. применение методов первичной обработки электронной информации. В случае сканированных изображений бумажных документов данный этап включает применение методов обработки изображений, повышающих точность распознавания.
4. Применение методов распознавания для преобразования информации, содержащейся в документе, в электронный вид для дальнейшего использования в системе электронного документооборота. Данный этап иногда включает в себя выделение некоторых полей (реквизитов) документа, для которых результат распознавания признан системой распознавания сомнительным или недостоверным, с последующей верификацией и коррекцией оператором.
5. Сохранение полученного электронного документа в базе данных и/или экспорт в удобный для электронной обработки формат, такой как XML, PDF, CSV и т.п.

Распределение потока документов на отдельные, независимо обрабатываемые пакеты представляет собой начальную стадию технологического процесса автоматического ввода документов и заключается в разбиении потока документов на части ограниченного размера и/или группировку документов по типу. Здесь и далее под типом документа подразумевается именованная совокупность его логической структуры (заголовок, множество полей (реквизитов) с определенными семантическими и синтаксическими свойствами) и структуры его представления на бумажном или ином физическом носителе.

Оцифровка документа является определяющим этапом для технологии автоматического ввода документов и представляет собой преобразование документа с физического носителя в электронный вид, удобный для дальнейшей обработки. К примеру, в случае оцифровки документов при помощи сканирования, для плоского (чаще всего – бумажного) документа строится его цифровое описание в виде цветного (многоканального), либо полутонового (одноканального) изображения с глубиной цвета и разрешением, которые регулируются в зависимости от технологических возможностей сканирующего устройства и от особенностей дальнейших алгоритмов обработки изображения документа. Другим примером оцифровки документа является его видео- или фотосъемка при

помощи камеры мобильного устройства, имеющая место в случае необходимости осуществлять эффективный ввод документов в нестационарном режиме. В этом случае электронным образом документа является его цифровая фотография либо видеопоток, содержащий упорядоченную последовательность кадров, на каждом из которых отображен документ или его часть.

Методы первичной обработки электронной информации, такие как обработка цифровых изображений, анализ и установление связей между информативными частями кадров в видеопотоке и т.д., применяются для облегчения задач выделения информативных областей цифрового образа документа и повышения точности распознавания. После первичной обработки в работу вступают методы определения логической структуры документа, выделения цифровых образов информационных полей (реквизитов) с последующим распознаванием. В зависимости от природы вводимых документов системы автоматического ввода документов используют методы оптического распознавания символов (Optical character recognition, OCR) [3], распознавания штрих-кодов (Barcode recognition, BCR) [4] и т.п. Методы оптического распознавания символов иногда подразделяют по функциональной направленности на методы распознавания печатных символов и печатного текста, рукопечатных символов, рукописных символов и рукописного текста, а также методы распознавания меток (к примеру, в анкетах с множественным выбором, избирательных бюллетенях и т.д.). В случае, если заранее известны синтаксические и/или семантические свойства полей (реквизитов) документа, после распознавания может производиться автоматическая коррекция результатов (к примеру, для коррекции результатов распознавания поля «Фамилия» может использоваться полный, либо неполный частотный словарь фамилий [5]. В некоторых системах автоматического ввода документов после того, как получен результат распознавания поля, производится анализ достоверности результата с последующей верификацией и коррекцией оператором [6; 7].

1.2 Мобильный документооборот

Начиная с 2000-х годов появляется широкий интерес к методам автоматического ввода документов с использованием мобильных устройств. Обусловлено

это быстро растущими вычислительными возможностями таких широко распространенных мобильных устройств, как «смартфоны» и портативные планшетные компьютеры, а также увеличивающимися техническими возможностями цифровых камер, установленных на этих устройствах. Интерес к системам электронного документооборота и, в частности, к методам автоматического ввода документов, применительно к мобильным устройствам также обусловлен развитием систем распространения мобильных приложений, как корпоративных, так и нацеленных на широкую публику. Согласно опросу пользователей мобильных устройств, который проводился в США в 2014 году компанией Radium One [8], 88.2% опрошенных пользуются своими смартфонами чаще, чем 10 раз в день, 35.5% – более 40 раз в день.

В корпоративном секторе повышается интерес к реализации делопроизводства (или его части) на основе мобильного документооборота – разновидности электронного документооборота, пользователи которого получают возможность производить операции с электронными документами при помощи различных мобильных устройств. Согласно опросу, который проводился компанией Litera Corp. в 2013-м году, 97% опрошенных профессиональных работников сферы бизнеса используют персональные, либо корпоративные мобильные устройства для хранения и обработки документов [9]. Естественным образом встает задача реализации систем автоматического ввода документов, использующих цифровые камеры мобильных устройств в качестве «сканирующего» устройства – оцифровка документа производится путем видео- или фотосъемки оригинала.

Среди обычных пользователей таких мобильных устройств, как смартфоны или планшетные компьютеры, возрастает интерес к приобретению товаров и услуг, совершая транзакции через интернет-сервисы, доступные с персональных мобильных устройств. Согласно ранее упомянутому опросу [8] 61% опрошенных пользователей смартфонов хотя бы раз совершали мобильную покупку в течение последних 6-ти месяцев. Согласно опросу 2014-го года, проводившемуся в 18-ти европейских стран, 77% опрошенных хотя бы один раз в жизни совершали мобильную покупку (против 72% в 2013-м году) [10]. В большинстве случаев заключение таких сделок подразумевает ввод данных некоторых документов (к примеру, документа, удостоверяющего личность, реквизиты банковской карты и т.д.), причем ввод этих данных зачастую требуется производить неоднократно, т.к. хранение этих данных в памяти мобильного устройства может привести

к утечке данных и их использованию злоумышленниками. Хранение чувствительных персональных данных на интернет-серверах строго ограничивается законодательством [11] и также, хоть и в меньшей степени, подвержено атакам со стороны мошенников. Это приводит к тому, что методы автоматического ввода документов, ориентированные на мобильные устройства, приобретают актуальность не только в корпоративной сфере, но и в сфере массовой электронной коммерции.

Еще одним двигателем, благодаря которому возрастает актуальность систем мобильного документооборота и мобильного распознавания документов, выступает роль комплекса процедур «Знай своего клиента» (англ. know your customer, KYC), согласно которому биржевым и банковским организациям а также другим финансовым институтам необходима точная идентификация клиента или контрагента для проведения финансовых операций. В рамках соответствия требованиям, собирательно относящихся к принципу «Знай своего клиента» ([12; 13]), клиентоориентированные финансовые организации вынуждены прибегать к идентификации пользователей и контрагентов при осуществлении каждой операции. Так как доля операций, осуществляемых удаленно при помощи мобильных устройств, растет, необходимость удаленной идентификации пользователей влечет к необходимости проводить удаленный анализ документов, в том числе документов, удостоверяющих личность.

Поскольку внедрение технологических, социальных и коммерческих процессов, основанных на использовании мобильных устройств и технологий, в условиях современного мира уже является обыденностью, системы автоматического ввода и анализа документов на мобильных устройствах продолжают вытеснять традиционные стационарные системы, и развитие технологий анализа документов с применением мобильных устройств и в условиях аппаратных ограничений, связанных с ними, является актуальной задачей.

1.3 Системы распознавания документов

Целью данного обзорного раздела является выделение основных этапов обработки изображений документов, характерных для систем автоматического ввода, и описание их особенностей.

1.3.1 Цифровой образ документа

Классические системы распознавания и автоматического ввода документов предполагают использование сканированного изображения документа в качестве его оцифрованного представления. Изображение в процессе оцифровки генерируется при помощи планшетного либо протяжного сканера, и характеризуется рядом особенностей: такое изображение, как правило, имеет высокое разрешение, поскольку разрешающая способность современных сканеров позволяют генерировать изображение с несколькими тысячами точек на дюйм. Освещение документа в подобных сканерах, как правило, равномерное, поскольку обеспечивается гомогенной искусственной подсветкой, и геометрический образ документа максимально соответствует оригиналу с точностью до небольших искажений в рамках расширенной группы движения.

Подавляющее большинство работ, связанных с автоматическим вводом и распознаванием документов на мобильных устройствах, рассматривают *фотографию* документа как его электронное представление и отмечают трудности, связанные с подготовкой образа документа к распознаванию и с самим распознаванием [14].

Изображения документов, получаемые с камеры мобильного устройства обладают гораздо более низким качеством, чем изображения, получаемые с традиционного цифрового сканера. В случае мобильных устройств на этапе подготовки изображения к распознаванию приходится сталкиваться с такими проблемами, как неравномерное освещение сцены, проективные искажения документа, нелинейные искажения документа (вызванные, к примеру, изгибом бумажного носителя), искажения, обусловленные движением камеры, зашумление, дефокусировка [15]. Все эти условия приводят к тому, что традиционные методы предварительной обработки изображения, применяемые в системах автоматического ввода документов с использованием цифровых сканеров не дают необходимого эффекта и появляется необходимость в специальных методах, позволяющих увеличить точность и надежность распознавания.

1.3.2 Поиск и локализация документа

Первичной задачей обработки изображения документа в системе распознавания является точный поиск документа на изображении. Как правило, данный этап также пересекается с задачей идентификации типа документа. В некоторых выделенных случаях данный этап опускается (в случае, когда тип документа полностью известен и изображение документа не имеет пространственных искажений ввиду специфики процесса оцифровки), однако в большинстве случаев этот этап необходим для дальнейшего анализа содержимого документа. Основными проблемами, с которыми приходится сталкиваться на этапе поиска документа на оцифрованном изображении, являются искажения изображения – как геометрические (наклоны, вращения, проективные искажения, нелинейные искажения), так и пиксельные (шумы оцифровки, яркостные искажения ввиду неравномерного освещения и т.п.).

Основные подходы к определению наклона документа на изображения можно разделить на две группы: глобальные и локальные [16]. Глобальные подходы анализируют признаки, вычисляемые по всему изображению, такие как гистограммы проекций объектов изображения на различные оси, прямые на границах областей изображения и т.п. Локальные подходы используют признаки, значимые только в ограниченных областях изображения, к примеру, общие оси соседних компонент связности текста. После этого на основе локально вычисленных оценках наклона документа принимается решение о глобальном значении оценки наклона. В работе [17] локальная оценка наклона вычисляется при помощи поиска направления с наибольшим количеством переходов от черного к белому и обратно в локальном окне бинаризованного изображения, содержащего текст. В работах [18; 19] описаны методы оценки локального наклона, также использующие геометрические свойства бинаризованного изображения текста. В работе [18] предлагается производить поиск цепочек компонент связности, которые соответствуют словам или частям слов или текстовых строк, при помощи метода наращивания регионов [20] и выбирать направление в окне согласно направлениям этих цепочек. В работе [19] предлагается метод, основанный на анализе направлений штрихов отдельных букв. В работах [21—23] угол наклона документа определяется при помощи преобразования Хафа [24], для поиска прямых на изображения документа. В работе [25] метод определения

наклона опирается на утверждение о том, что для компонент документа, содержащих текст и другие типичные элементы (штрих-коды, таблицы и т. п.) основная ось охватывающего прямоугольника минимальной площади совпадает с направлением компоненты. Предлагаемый алгоритм поиск прямоугольника минимальной площади для отдельно взятой компоненты достаточно эффективен и опирается на метод следования вдоль границы компоненты, описанный в [26].

Достаточно большое внимание уделяется задаче проективного исправления изображения документа, полученного с камеры мобильного устройства [27—31]. Большинство методов исправления перспективных (проективных) искажений изображения основаны на детектировании исчезающей точки перспективы (vanishing point). Эти методы включают в себя поиск линий и точек пересечения этих линий на изображении, либо текстурный и частотный анализ составляющих изображения [32]. Для поиска исчезающей точки перспективы или линий документа в свою очередь используется поиск границ документа (если на изображении присутствует весь документ, либо целиком одна из его страниц), либо поиск линий текстовых строк. Поиск линий на изображениях – задача хорошо описанная в литературе [33], для ее решения предложено множество методов, основанных на применении алгоритма RANSAC [34], поиске прямых методом наименьших квадратов [35], а также быстрого преобразования Хафа [24].

Некоторые методы основаны на анализе направлений текстовых линий (эти методы необходимы в случаях, когда границы документа не попадают в кадр). В работе [36] предлагается метод, основанный на анализе компонент связности (символов) на изображении. Ректификация (исправление проективного искажения) в предложенном методе производится отдельно для каждой текстовой строки на основе аппроксимации ее базовых линий.

Помимо методов, анализирующих отдельные элементы документов с целью ректификации изображения (т.е. частично сводящих задачу локализации документа к определению параметров частных геометрических искажений), появляется и другой класс методов, связанных с непосредственно локализацией образа страницы документа целиком. Одним из подходов к локализации страницы документа по его визуальному представлению является использование обобщение метода Виолы и Джонса [37] как решающего дерева сильных классификаторов [38; 39] для детектирования образа страницы целиком в условиях ограниченных перспективных искажений. Однако наиболее широко используе-

мым подходом здесь стоит отметить подход поиска опорных точек (keypoints или feature points). Мотивацией к применению данного подхода для решения этой задачи является их устойчивость к различного рода помехам и искажениям изображения, которые могут существенно понизить точность работы алгоритмов структурного и геометрического анализа [40].

Описанные подходы, сами по себе успешно используемые для анализа изображений документов на отдельных изображениях, вообще говоря не обобщены на случай множества изображений одного и того же документа, и, таким образом, относятся к статическим системам распознавания объектов. Рассматривая видеопоток как цифровой образ распознаваемого объекта данные подходы нуждаются в соответствующей адаптации.

1.3.3 Сегментация изображения документа

Следующим этапом после поиска и локализации документа на изображении является сегментация (уже ректифицированного) изображения документа на составные части. Постановка данной задачи зависит от структуры документа и от дальнейших методов обработки, однако из них можно выделить две крупные задачи, которые редко пересекаются и в большинстве систем распознавания представляют собой два независимых этапа:

1. Сегментация изображения документа на отдельные фрагменты – текстовые блоки, текстовые поля, строки, параграфы, фигуры и формулы, печати и т.п.;
2. Сегментация изображения текстовой строки или текстового поля на отдельные символы/графемы.

Поскольку сегментация изображения документа на информационные фрагменты в значительной степени зависит как от структуры самого документа, так и от специфики конкретных систем распознавания, в литературе наблюдается широкий спектр методов и подходов, предлагаемых для решения этой задачи. Для сегментации изображения на крупные гомогенные информационные блоки предлагаются методы комбинаторного анализа прямолинейных элементов, детектируемых при помощи морфологического анализа изображения [41], декомпозиция изображения документа при помощи оптимального накладыва-

ния множества Гауссовых ядер [42], либо использование полностью сверточных искусственных нейронных сетей (ИНС) для выделения пикселей, соответствующих тому или иному сегментируемому фрагменту [43]. Также развиваются структурные методы поиска текстовых строк на основе поиска путей в яркостном графе [44] и методы, частично использующие техники машинного обучения для детектирования границ текстовых строк [45].

Отдельно стоит выделить подход к поиску текстовых строк и отдельных слов, наиболее широко обсуждаемый в литературе в течении нескольких последних лет, а именно подход «word spotting», или «text in the wild». Этот подход образовался в рамках более общей задачи поиска текста на произвольных естественных изображениях. В рамках решения более общей задачи появился ряд методов, позволяющий выделять участки изображения, содержащие символы, графемы, слова и текстовые строки целиком, в условиях пиксельных и геометрических искажений. Методы включают как структурный анализ участков изображения с последующим комбинаторным выбором участков, обладающих признаками текста [46—48], использование техник машинного обучения локальных признаков текста [48—50], а также глобальное использование глубоких сверточных ИНС для выделения пикселей текста на произвольном изображении [49; 51—53].

Одной из самых сложных задач распознавания текста, и, в частности, систем распознавания документов, остается задача сегментация изображения текстовой строки на отдельные символы [54; 55]. Процедура сегментация на отдельные символы, помимо общих искажений, которые претерпевает изображение документа (влекущих за собой «склейки» соседних символов и графем и другие проблемы), затруднена большим многообразием шрифтов и гарнитур, используемых для печати текстовых полей документов. В традиционных системах распознавания документов для задачи сегментации текстовой строки на отдельные символы использовалась бинаризация изображения текстовой строки (либо глобальная бинаризация всего изображения документа) с последующим анализом «черных» компонент связности. Однако в условиях искажений, характерных для изображений, полученных при помощи камеры мобильного устройства, данный метод перестал быть актуальным, поскольку подбор универсальных параметров бинаризации, при которых минимизируется количество «склеек» и «разрезов», чрезвычайно затруднен. Наиболее широко используемые подходы к этой задаче без бинаризации включают анализ проекции изображе-

ния текстовой строки на горизонтальную ось, расстановку предварительных разрезов строки и применение динамического программирования с использованием значимости разрезов и оценок достоверности распознавания символов между разрезами в качестве составных частей финальной метрики «качества» сегментации [56; 57]. Другие методы используют техники машинного обучения для обучаемой расстановки разрезов между символами/графемами: использование сверточных ИНС и рекуррентных LSTM-сетей («long short-term memory networks») позволяет уменьшить количество эвристических параметров алгоритма сегментации и, в то же время, увеличить устойчивость к искажениям [56].

Также стоит отметить, что популярность среди исследователей набирают методы распознавания текстовых строк, полностью исключающих явную сегментацию на отдельные символы. В данных подходах предполагается объединение глубоких сверточных ИНС с LSTM-сетями для обучения алгоритма распознавания целиком слов, без промежуточных этапов [58; 59]. Такие методы позволяют значительно повысить устойчивость алгоритма к искажениям входного изображения, и позволяют расширить домен применения алгоритма (к примеру, на задачу распознавания рукописных строк, текста, напечатанного сложно сегментируемым языком, таким как персидский или арабский, или строк исторических документов со сложно сегментируемыми каллиграфическими шрифтами). Однако сложность обучения нейросетевых моделей, решающих настолько общую задачу, и высокая трудоемкость алгоритма распознавания каждой строки пока не позволяют использовать данные подходы в практических реализациях систем распознавания, особенно принимая во внимание аппаратные ограничения систем распознавания изображения на мобильных устройствах.

1.3.4 Распознавание одиночных символов

Оптическое распознавание объектов вообще и, в частности, печатных или рукописных символов, является одной из наиболее важных задач компьютерного зрения. За последние несколько десятилетий возник значительный спектр методов, применяемых для решения задачи распознавания образов различных объектов. В классическом представлении системы распознавания по свойствам

информации, используемой в процессе распознавания, подразделяются на системы без обучения, системы на основе обучения, детерминированные, вероятностные, логические, структурные и комбинированные [60].

Искусственные нейронные сети (ИНС), механизм, изначально направленный на моделирование биологических систем, на данный момент являются одним из наиболее эффективных механизмов распознавания образов и, в частности, распознавания символов. В ряде отдельных задач данный метод показывает себя способным конкурировать с системой восприятия образов человеком [61]. В качестве мотивации искусственных нейронных сетей использовалась упрощенная модель мозга [62]. Современные исследования и разработки, ведущиеся в области машинного обучения, продолжают развивать архитектуры и методики обучения ИНС, предназначенные для решения разнообразных задач. Фундаментальным прорывом в области анализа изображений при помощи ИНС стало предложение сверточных ИНС.

Сверточная ИНС была впервые предложена Яном Лекуном [63], и ее архитектура предполагала два сверточных слоя, два прореживающих слоя и нескольких полносвязных слоев для формирования результата. Такая архитектура позволила сделать отклик ИНС инвариантным к координатному сдвигу исходного сигнала, а обработку признаков — одинаковой для разных локальных областей входного изображения. На основе оригинальной идеи сверточных ИНС построено множество нейросетевых архитектур, приспособленных для решения конкретных задач, к примеру, AlexNet [64], в которой также предложен распределенный способ обучения на нескольких графических сопроцессорах; ZFNet [65], в которой за счет увеличения размера средних сверточных слоев был улучшен подбор гиперпараметров, а также был предложен вариант визуализации сети; VGGNet [66], в рамках которой было показано, что глубина сети является важной компонентой качества в задаче классификации масштабных изображений; и другие архитектуры [67—69].

Использование сверточных ИНС в настоящий момент является наиболее точным методом распознавания изображений, и в ряде отдельных задач этот метод показывает результаты, способные конкурировать с человеком [61]. Тем не менее сверточные ИНС могут показывать неустойчивый результат при минимальных изменениях входного изображения [70; 71], даже если эти изменения касались всего лишь одного пикселя [72]. Еще одной проблемой сверточных ИНС, связанной с предыдущей, является их чрезмерная «самоуверенность»

(overconfidence) — оценки, выдаваемые сверточными ИНС на некорректно распознанных изображениях могут быть неотличимы от оценок, содержащихся в правильных результатах [73].

Применение сверточных ИНС в мобильных системах распознавания документов накладывает дополнительные ограничения, связанные с высокой, по сравнению с некоторыми другими методами, вычислительной трудоемкостью их использования, которая значительно возрастает при увеличении размерности входа. Для того, чтобы использовать сверточные ИНС в системах распознавания документов, использующих несколько изображений (к примеру, распознавания объектов в видеопотоке), необходима разработка высокоэффективных методов комбинации полученных результатов распознавания отдельных изображений.

1.3.5 Пост-процессинг и языковые модели

В задаче распознавания текстовых полей документов будет ошибкой ограничиться конкатенацией независимых результатов классификации символов, поскольку для многих полей известна синтаксическая и семантическая структура. Совокупность представлений о допустимых символах в определенных знакоместах текстового поля, о их взаимозависимостях внутри поля и о зависимостях между значениями разных текстовых полей одного и того же документа, будем называть контекстом текстового поля. Построение алгоритмов контекстно-зависимого уточнения результатов распознавания как правило зависит от специфики задачи, и в этой области постоянно предлагаются новые подходы к решению конкретных прикладных задач [74; 75]. Таким образом статистическая коррекция («пост-обработка» или «пост-процессинг») результатов распознавания является одним из важнейших компонентов современных систем оптического распознавания документов.

Контекст поля документа, как правило, включает в себя следующие компоненты:

1. *синтаксис:* правила, регулирующие структуру текстового представления поля;

2. *семантика поля:* правила, основывающиеся на смысловой интерпретации поля или его составных частей;
3. *семантика связей:* правила, основывающиеся на структурной и смысловой связи поля с другими полями документа.

Существует множество алгоритмов статистической пост-обработки результатов распознавания, отличающихся в используемых языковых моделях распознаваемого объекта, в алгоритмах непосредственного исправления результата распознавания и в областях применимости. Среди наиболее известных и широко используемых методов можно выделить: методы, опирающиеся на скрытые марковские модели (Hidden Markov Models, HMM) [76; 77], конечные автоматы, N-граммные и словарные методы [5; 78], а также механизмы, использующие взвешенные конечные преобразователи (Weighted Finite-State Transducers, WFST) [79].

Располагая информацией о семантической и синтаксической структуре документа и распознаваемого поля, можно построить специализированный алгоритм пост-обработки для каждого конкретного поля. Однако, принимая во внимание необходимость поддержки и развития систем распознавания и сложность их разработки, особый интерес представляют методы и инструменты, позволяющие с минимальными усилиями (со стороны разработчиков системы распознавания) построить достаточно хороший алгоритм пост-обработки, который бы работал с обширным классом документов и полей. Методика настройки и поддержки такого алгоритма была бы унифицирована, а изменяемым компонентом структуры алгоритма были бы только семантика и синтаксис обрабатываемого поля. Достаточно общая модель, позволяющая построить универсальный алгоритм пост-обработки результатов распознавания, описана в работе [79]. Модель опирается на структуру данных взвешенных конечных преобразователей (Weighted Finite-State Transducers, WFST).

Преимуществами данного подхода является его общность и гибкость. Модель ошибок, к примеру, может быть без труда расширена таким образом, чтобы учесть удаления и добавления символов (для этого всего лишь стоит добавить в модель ошибок переходы с пустым выходным или входным символом соответственно). Однако у такой модели есть и существенные недостатки. Во-первых, языковая модель здесь должна быть представлена в виде конечного взвешенного конечного преобразователя. Для сложных языков такой автомат может получиться довольно громоздким, и в случае изменения или уточнения языко-

вой модели будет необходимо его перестроение. Также необходимо заметить, что композиция трех преобразователей в качестве результата имеет, как правило, еще более громоздкий преобразователь, а эта композиция вычисляется каждый раз при запуске пост-обработки одного результата распознавания. Ввиду громоздкости композиции, поиск оптимального пути на практике приходится выполнять эвристическими методами [79], такими как A*-search [80].

1.3.6 Оценка достоверности распознавания

При оценке качества систем и алгоритмов распознавания используются такие понятия как *точность* и *уверенность*. Точность системы распознавания отражает вероятность (оценку вероятности) правильного распознавания объекта. Так как истинное значение вероятности правильного распознавания в общем случае определить нельзя, точность систем распознавания обычно оценивается вычислением апостериорной вероятности правильного распознавания на некотором заранее заданном пакете входных данных. Такой «референтный» пакет данных часто является основой для постановки задачи распознавания: при заданном референтном пакете данных задача распознавания может быть сформулирована как задача максимизации апостериорной вероятности правильного ответа для объектов из этого пакета данных.

Еще одним важным показателем качества работы системы распознавания объектов является ее уверенность (в некоторых источниках также используется схожий термин «надежность распознавания» или «достоверность распознавания», для того, чтобы уменьшить возможные коллизии терминов в дальнейшем будет использоваться термин «оценка достоверности»). В системах распознавания объектов различного уровня сложности часто возникают задачи моделирования взаимодействия системы с пользователем, а также взаимодействия подсистем между собой. Одним из важнейших аспектов таких моделей взаимодействия является реакция системы на ошибки распознавания. Оценка достоверности системы распознавания отражает ее способность априорно оценить степень точности собственного результата.

Задача оценки достоверности результатов распознавания в рамках единой модели взаимодействия системы автоматической обработки документов была

рассмотрена в работе [6]. В указанной работе автор описал схему построения функции оценки эффективности правила отбраковки, исходя из вычисленных апостериорных вероятностей ошибок классификатора и стоимостей этих ошибок, задаваемых извне. Также были рассмотрены простейшие правила отбраковки результата распознавания (правило первой альтернативы, правило двух альтернатив) и сформулированы общие принципы построения комплексных правил отбраковки. В данном разделе будет представлена попытка реализации общего метода построения функции достоверности и соответствующего ей решающего правила, основываясь на заранее определенных предикторах (признаках достоверности), вычисляемых по вектору альтернатив классификатора или по исходному изображению, минимизируя при этом штрафной функционал.

Исследование методов оптимальной отбраковки результатов распознавания продолжаются как в сторону построения методов отбраковки для отдельных классификаторов, так и в сторону моделирование оптимальных схем отбраковки в системах с несколькими классификаторами, либо в системах человеко-машинного взаимодействия [81].

1.3.7 Использование множества входных изображений

Для задачи распознавания одиночного объекта в видеопотоке при существующем решении задачи распознавания одиночного объекта на одном кадре можно руководствоваться двумя основными подходами:

1. Производить первоначальное совмещение изображений объектов, после чего производить распознавание совмещенного объекта;
2. Производить распознавание изображения объекта на каждом отдельном кадре, после чего решать задачу комбинирования (интеграции) результатов распознавания.

Системы, предполагающие использование мобильных устройств в задачах автоматического ввода документов, располагают рядом преимуществ как с точки зрения пользователя, так и с точки зрения постановки задачи распознавания образов. В системе автоматического ввода документов, которая использует цифровые камеры мобильных устройств в качестве сканирующего устройства, оцифровка документа производится путем видео- или фотосъемки оригинала.

В случае видеосъемки документа цифровым представлением документа является не единственное изображение, а последовательность кадров, содержащих образ одного и того же документа (или его фрагмента), что обеспечивает возможность производить многократное распознавание одного и того же объекта в видеопотоке в реальном времени, увеличивая тем самым точность и надежность распознавания.

При этом использование мобильных устройств в задачах распознавания и автоматического ввода документов сопряжено с рядом проблем: помимо низких вычислительных мощностей мобильных устройств, к этим проблемам следует отнести широкий спектр искажений, возникающих вследствие особенностей оптической схемы малоформатных камер мобильных устройств [82], а также вследствие особенностей процесса съемки документа при помощи мобильного устройства [15; 83]. Часть ошибок распознавания, связанных с этими проблемами, можно решить при помощи многократного распознавания одного и того же объекта, при этом естественным образом возникает задача выбора оптимальной стратегии комбинирования результатов покадрового распознавания.

В зависимости от используемой модели результата классификации объекта и от интерпретации оценок классификатора рассматриваются различные методы комбинирования. Для модели результата классификации в виде пары $\langle\sigma,q\rangle$, где σ – метка класса, а q – показатель уверенности классификатора (без альтернатив), используется схема голосования с выбором лучшего ответа с максимизацией некоторого количественного критерия. Обобщенный критерий приведен в работе [84]:

$$Score\left(\langle\sigma,q\rangle\right) = \alpha \cdot Freq(\sigma) + (1-\alpha)\cdot q, \tag{1.1}$$

где $Freq(\sigma)$ — частота класса σ среди ответов комбинируемых классификаторов, а α — обучаемый параметр алгоритма. Распространенным методом, основывающимся на голосовании классификаторов и применяющийся в системах распознавания речи при помощи комбинирования нескольких алгоритмов, а также для комбинирования результатов распознавания текстовых строк [85; 86], является метод ROVER (Recognizer Output Voting Error Reduction) [87].

На рисунке 1.1 представлены примеры текстовых полей в видеопотоке, и в Таблице 1 приведены соответствующие покадровые результаты распознавания, а также интегрированные методом ROVER [87] результаты и их изменение во времени.

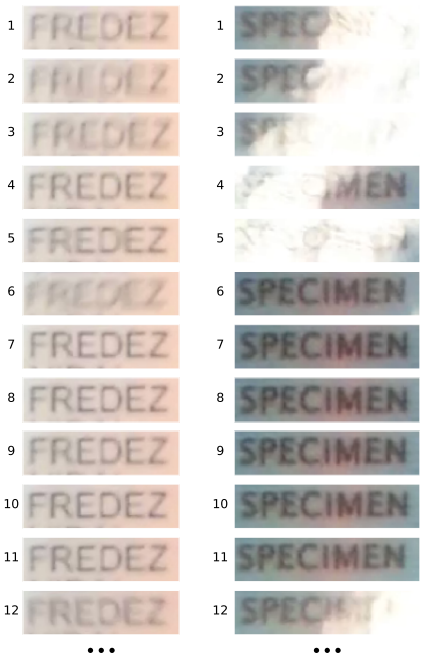


Рисунок 1.1 — Примеры изображений текстовых полей в видеопотоке из пакета данных MIDV-500 [88] (видеоклипы TS07, поле 1, и HA10, поле 2)

Таблица 1 — Примеры покадровых и интегрированных результатов распознавания текстовых полей. Верные результаты выделены.

#	Покадровый	Интегрированный	Покадровый	Интегрированный
1	{REOCZ	{REOCZ	-	-
2	HUD‘‘	{REOCZ	I	-
3	MON	EOCZ	W	-
4	FREDEZ	REOCZ	A-	-
5	‘REZIZ	REOCZ	3" J" .l	-
6	MM	REOCZ	‘‘(MEN	-
7	FREDEZ	FREOEZ	SPECIMEN	E-
8	FREDEZ	FREDEZ	SPECIMEN	MEN
9	FREDEZ	FREDEZ	SPECIMEN	CIMEN
10	FREDEZ	FREDEZ	SPICIUEN	SPCIMEN
11	FREDEZ	FREDEZ	SPECIMEN	SPECIMEN
12	RREDEZ	FREDEZ	.-	SPECIMEN
...	...	...	...	...

В случае более общей модели результата классификации, с интерпретацией оценок принадлежности q_k в рамках Байесовской модели (т.е. оценка принадлежности q_k есть апостериорная оценка условной вероятности принадлежности образа x_i к классу σ_k), описываются различные правила комбинации [89; 90]. Базовые правила описаны в работе [90], которая признается фундаментальным

трудом, связанным с задачей объединения результатов различных классификаторов в рамках Байесовской модели. Опишем эти правила, но в применении к задаче объединения результатов классификации нескольких образов одним классификатором:

1. *Правило произведения:*

$$\mathrm{Prod}(X)(\sigma) = P(\sigma|X) = \frac{1}{P(\sigma)^{N-1}} \prod_{i=1}^{N} C_\Sigma(X_i)(\sigma); \tag{1.2}$$

2. *Правило суммы:*

$$\mathrm{Sum}(X)(\sigma) = \frac{1}{N} \sum_{i=1}^{N} C_\Sigma(X_i)(\sigma); \tag{1.3}$$

3. *Правило минимума:*

$$\mathrm{Min}(X)(\sigma) = \left(\min_{i=1}^{N} C_\Sigma(x_i)(\sigma) \right) \cdot \left(\sum_{k=1}^{K} \min_{i=1}^{N} C_\Sigma(x_i)(\sigma_k) \right)^{-1}; \tag{1.4}$$

4. *Правило максимума:*

$$\mathrm{Max}(X)(\sigma) = \left(\max_{i=1}^{N} C_\Sigma(x_i)(\sigma) \right) \cdot \left(\sum_{k=1}^{K} \max_{i=1}^{N} C_\Sigma(x_i)(\sigma_k) \right)^{-1}; \tag{1.5}$$

5. *Правило медианы:*

$$\mathrm{Med}(X)(\sigma) = \left(\operatorname*{med}_{i=1}^{N} C_\Sigma(x_i)(\sigma) \right) \cdot \left(\sum_{k=1}^{K} \operatorname*{med}_{i=1}^{N} C_\Sigma(x_i)(\sigma_k) \right)^{-1}. \tag{1.6}$$

В случае с интерпретацией оценок q_k как нечетких свидетельств принадлежности к классам либо абстрактных показателей уверенности, используются методы комбинирования основанные на теории Демпстера-Шафера [91; 92]. Также в работах, затрагивающих гетерогенные методы объединения результатов классификаторов, рассматриваются стратегии взвешивания уровней значимости классификаторов [93], методы обучения правил комбинирования, учитывающие статистические особенности объединяемых классификаторов [94—96] и методы, не привязанные к статистическим особенностям классификаторов, но использующие аппарат мультимножеств для построения модели групповой классификации объектов [97; 98].

Несмотря на большое количество описанных методов комбинирования результатов распознавания, полученных при помощи различных методов, подходам к комбинированию результатов распознавания различных изображений одного и того же объекта в литературе уделяется недостаточного внимания. При этом эта задача является одной из ключевых задач при построении систем распознавания объектов в видеопотоке.

Помимо задачи комбинирования результатов распознавания объектов на отдельных изображениях, при распознавании видеопотоке возникает также новая задача – задача останова процесса распознавания. Задача останова процесса распознавания является особенно актуальной применительно к системам компьютерного зрения и распознавания документов, оперирующих в реальном времени на мобильных устройствах [99—101], в которых время, необходимое для получения итогового результата настолько же важно, как и точность этого результата.

С точки зрения системной композиции процесс распознавания объектов в видеопотоке, при котором отдельные кадры распознаются независимо и комбинируются в единый результат, может рассматриваться как «anytime»-алгоритм (алгоритм с отсечением по времени: итерационный вычислительный алгоритм, который способен выдать наилучшее на данный момент решение в любое время, если процесс вычислений не доводится до естественного останова) [102]. Можно считать, что среди свойств «anytime»-алгоритмов процесс распознавания в видеопотоке обладает свойством *возможности прерывания* (interruptibility) (т.е. процесс может быть остановлен после обработки любого кадра и текущий интегрированный результат может быть принят за итоговый) и свойством *монотонности* (monotonicity) (т.е. качество интегрированных результатов в среднем не ухудшается). При этом процесс распознавания объекта в видеопотоке может не обладать свойством *определимого качества* (recognizable quality), т.е. качество текущего результата может быть неизвестно и невычислимо в момент исполнения алгоритма. Стоит отметить, что существуют другие примеры «anytime»-алгоритмов в распознавании и компьютерном зрении, к примеру, алгоритмы, способные выдать частичный результат: сначала для объектов, наиболее хорошо поддающихся распознаванию, и прогрессируя к более сложным объектам [103].

Задача останова в ее более общей постановке хорошо рассмотрена в литературе по математической статистике и теории принятия решений [104—107].

Изучены такие варианты задачи останова, как задача о разборчивой невесте [108], задача о продаже дома [109], задача оптимизации среднего (также известна как задача S_n/n^1) [110] и другие.

Одна из хорошо изученных задач останова, задача о вычитке [106; 111], может быть рассмотрена как наиболее близкая по содержанию к задаче останова процесса распознавания объекта в видеопотоке. Задача формулируется следующим образом: некоторая рукопись была оцифрована с количеством ошибок M. Для оцифрованной версии рукописи может быть проведена серия вычиток, каждая i-я из них исправляет X_i ошибок и обладает фиксированной стоимостью c. Каждая ошибка, присутствующая в финальной версии текста также привносит некоторый штраф. Задача после i-й вычитки состоит в принятии решения о том, что текущую версию оцифрованной рукописи следует подать на публикацию, либо процесс вычиток следует продолжить, с целью минимизации суммарной ожидаемой стоимости. Для этой задачи было предложено несколько вариантов решения [111; 112], опирающихся на различные предположения о распределениях M и X_i. У этой задачи также есть другие варианты постановки, такие, как задача о автоматическом тестировании программного обеспечения [113].

Между задачами о вычитке и задачей останова процесса распознавания в видеопотоке можно заметить несколько сходств: распознавание объекта производится на нескольких кадрах, и для того, чтобы получить очередной результат распознавания, должен быть уплачен некоторый штраф (выраженный, к примеру, во времени, которое необходимо для пред-обработки очередного кадра и проведения очередной итерации алгоритма распознавания объекта). Исходя из предположения, что на каждом шаге процесса определен некоторый аккумулированный результат распознавания, после обработки каждого кадра должно быть принято решения либо заплатить стоимость очередного наблюдения и продолжить процесс распознавания в надежде, что результат будет улучшен, или остановить процесс и вывести текущий результат. В данном случае ожидаемый убыток может быть представлен в виде линейной комбинации ожидаемого количества обработанных кадров и расстояния между ожидаемым результатом распознавания объекта до истинного значения (в терминах некоторой заранее определенной метрики).

Существуют также и важные отличия между этими задачами, на которые следует обратить внимание:

1. В большинстве формулировок задачи о вычитке предполагается, что X_i принимает неотрицательные значения, т.е. каждая вычитка либо исправляет некоторое количество ошибок, либо, по крайней мере, не привносит новых. В задаче распознавания объекта нет гарантии того, что результат распознавания на следующем кадре, ровно как и интегрированный результат после обработки следующего кадра, будет всегда ближе к истинному значению. С другой стороны, можно предположить, что алгоритмы интеграции результатов распознавания (см. главу 3) обычно конструируются таким образом, чтобы результаты распознавания множества версий одного и того же объекта в среднем обладают большей точностью, чем результат распознавания одного изображения объекта.
2. Решения задачи о вычитке или ее вариаций, как правило, опираются на возможность лица, принимающего решение, оценить стоимость, которая будет уплачена в случае принятия решения об останове, либо оценить разницу в стоимостях останова на соседних этапах, поскольку наблюдаемое значение X_i имеет прямой вклад в функцию штрафа. Напротив, в случае процесса распознавания объекта в видеопотоке, расстояние от вновь полученного результата до истинного значения может только оцениваться, опираясь на методы оценки уверенности результатов распознавания. Подобные оценки могут обладать свойством чрезмерной уверенности [114; 115], либо вовсе быть недоступны лицу, принимающему решение.

Подробное обсуждение задачи о вычитке представлено в работе [111] и показано, что оптимальное правило останова для данной задачи может быть быть построено с использованием понятие о *монотонных задачах останова*. Теория монотонных задач останова подробно описана в [116] и [106] и также будет использоваться в данной работе для построения метода останова процесса распознавания объекта в видеопотоке.

1.4 Выводы по аналитической части

Развитие систем автоматического ввода документов и систем оптического распознавания объектов обусловлено развитием большого количества дисциплин, таких как обработка изображений, компьютерное зрение, машинное обучение, проектирование систем программного обеспечения, а также развитием вычислительной техники, в том числе мобильных вычислительных устройств. Каждая из этих дисциплин оказывает влияние на эффективность отдельных компонент и подсистем, а тренды развития отдельных дисциплин определяют спектр алгоритмов, применяемых для решения частных задач, возникающих в сфере автоматического анализа и распознавания документов.

В литературе значительное внимание уделяется вопросам систем обработки изображений документов, алгоритмам поиска документов и их фрагментов на изображениях, методам сегментации изображений документов, детектированию и классификации различных объектов и реквизитов документа. Живой интерес к разработке высокоточных алгоритмов, решающих эти задачи, наличие публикаций в течении нескольких последних лет в крупных журналах и на профильных конференциях свидетельствует об актуальности данной темы. При этом внимание уделяется как структурным методам анализа изображений и их фрагментов, так и методам основанным целиком на машинном обучении. В особенности следует отметить тренд последних 3–4 лет к применению методов глубокого обучения для решения сложно поставленных задач, таких как целиковый поиск текстовых фрагментов на естественных изображениях и распознавания строки текста целиком минуя отдельно выделенный этап сегментации на отдельные символы. Однако стоит отметить, что несмотря на то, что данные методы показывают достаточно приемлемую точность распознавания и устойчивость к шумам входных данных, применение такого рода алгоритмов в промышленных системах распознавания и автоматического ввода документов затруднено их высокой трудоемкостью. Высокие вычислительные требования подобного рода методов делает их применение особенно неоправданным в контексте мобильных вычислительных устройств.

Распознавание объектов в видеопотоке является относительно новой темой в контексте систем распознавания и автоматического ввода документов. Существуют описанные в литературе теоретические подходы к групповой клас-

сификации многопризнаковых объектов и отдельные работы по попыткам применения многократного распознавания и к комбинации результатов выходов классификаторов, однако подробное изучение методов повышения точности распознавания объектов за счет многократного распознавания остается актуальной и малоизученной темой. При этом рассматривая видеопоток, а не одиночное изображение, как цифровой образ распознаваемого объекта, задача комбинирования результатов распознавания различных наблюдений одного и того же объекта является ключевой.

Отдельной актуальной темой является задача останова процесса распознавания объектов в видеопотоке, имеющая критическое значение применительно к системам компьютерного зрения и распознавания документов, оперирующих в реальном времени на мобильных устройствах. Хотя в литературе известны и теоретически проработаны большое количество различных задач останова, необходимо уделить внимание способам применения этой теории в системах распознавания объектов в видеопотоке.

1.5 Задачи диссертационной работы

На основе проведенного анализа основных принципов современных систем автоматического ввода документов и систем оптического распознавания документов, сформулированы следующие актуальные задачи для диссертационной работы:

1. Построить математическую модель системы распознавания объекта в видеопотоке, позволяющую исследовать качественные характеристики результата и время, необходимое для его получения;
2. Исследовать влияние характеристик входных данных на выбор оптимальной стратегии комбинирования результатов распознавания;
3. Разработать алгоритм комбинирования результатов оптического распознавания строкового объекта и провести анализ его характеристик;
4. Разработать метод останова процесса распознавания объекта в видеопотоке в рамках построенной математической модели системы;
5. Разработать алгоритм останова процесса распознавания строкового объекта и провести экспериментальны анализ его характеристик.

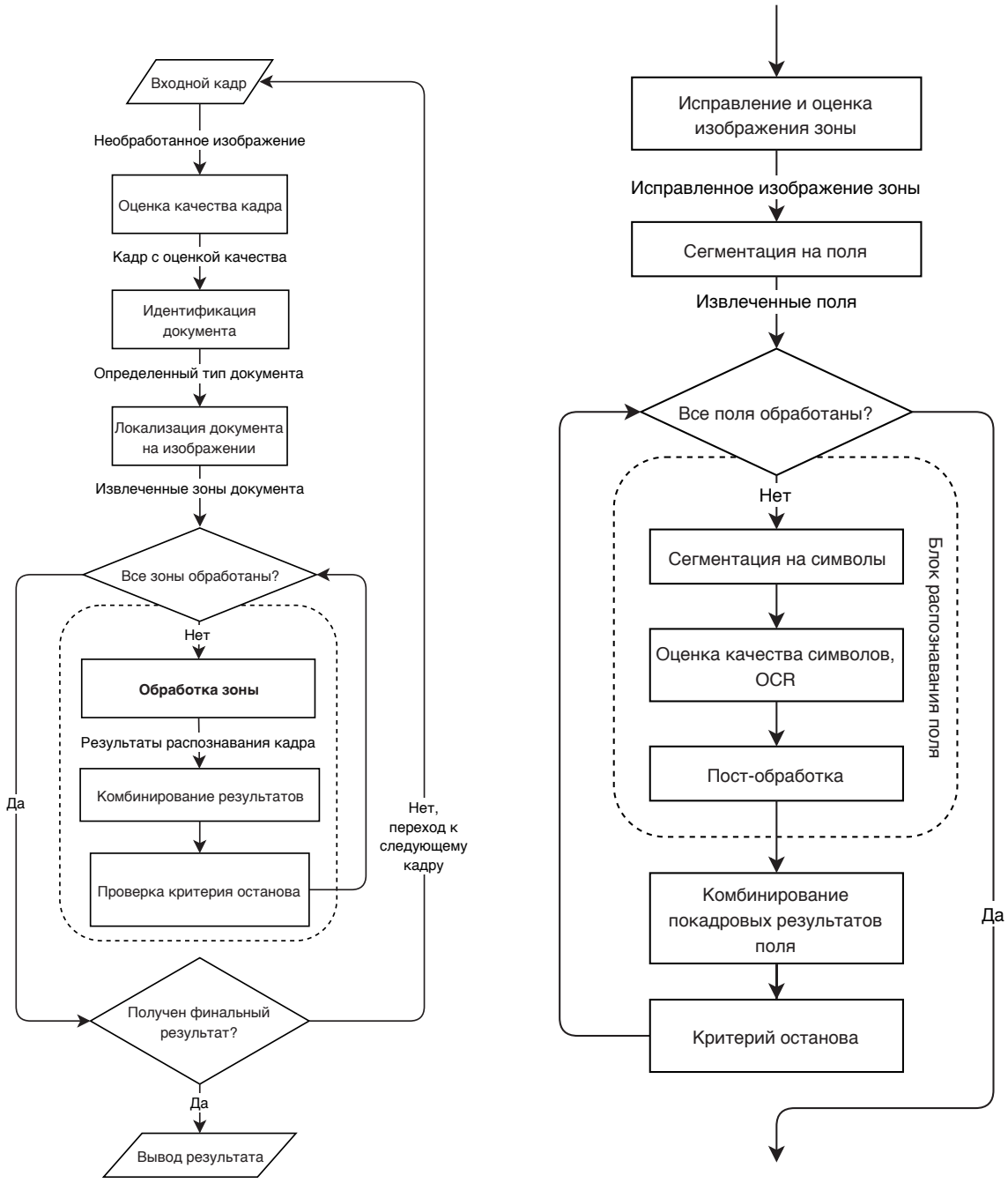
Глава 2. Модель системы распознавания объектов в видеопотоке мобильного устройства

2.1 Введение

Внедрение технологических, социальных и коммерческих процессов, основанных на использовании мобильных устройств и технологий, в условиях современного мира уже является обыденностью. Системы технического зрения с использованием мобильных технологий, к примеру, системы автоматического ввода и анализа документов на мобильных устройствах продолжают вытеснять традиционные стационарные системы, и развитие технологий технического зрения с применением мобильных устройств и в условиях аппаратных ограничений, связанных с ними, становится все более актуальной задачей.

Классические системы распознавания и автоматического ввода предполагают использование сканированного изображения или фотографии объекта в качестве его оцифрованного представления. При использовании мобильных устройств для оцифровки образов распознаваемых объектов возникает дополнительная возможность использовать видеопоток цифровой камеры помимо отдельных фотографий или кадров. Процесс съемки фотографии объекта при помощи современных мобильных устройств предполагает этап «наведения» оператором объектива камеры на объект с отображением кадров видеопотока на экране устройства в реальном времени для контроля оператора. В случае, если обработка изображения производится с одного изображения, информация, которая содержится в захваченных предварительных кадрах используется лишь косвенно (оператором). При рассмотрении цельного видеопотока в качестве цифрового образа объекта появляется возможность использовать гораздо больше визуальной информации [99]. Схема рассматриваемых систем автоматического ввода документов в видеопотоке представлена на рисунке 2.1.

Использование видеопотока позволяет решать задачи, недоступные для решения при анализе одиночной фотографии. Внешние условия съемки могут привести к тому, что распознаваемый объект сильно искажен на одиночном изображении [15]. Примером является блик от протяженного источника света, проявляющийся на глянцевой поверхности плоского объекта (см. рис. 2.2).



а) Общая схема б) Схема обработки зон

Рисунок 2.1 — Схема обработки кадра в системе распознавания документов в видеопотоке. Слева (а) – общая схема, справа (б) – схема блока обработки зоны документа (обведен пунктиром на общей схеме).

Поскольку в видеопотоке геометрическое положение снимаемого объекта, как правило, меняется между кадрами, блик также «сдвигается», что позволяет получить информацию о скрываемом объекте на другом кадре видеопотока. Существуют также важный класс объектов, детектирование и распознавание которых невозможно на одиночных снимках — к примеру, голографические эле-

Рисунок 2.2 — Процесс съемки идентификационного документа при помощи мобильного устройства (в качестве документа используется макет идентификационной карты Германии).

менты защиты, которые на единичных изображениях могут быть неотличимы от бликов или рисунков [15].

В таких условиях возникает задача выбора оптимальной стратегии комбинирования результатов покадрового распознавания. Данная задача в литературе практически не описана, и наиболее близкий спектр методов касается задачи комбинирования результатов распознавания одного и того же объекта, но разными классификаторами [85; 89; 90]. Помимо базовых стратегий объединения оценок в работах, затрагивающих гетерогенные методы объединения результатов классификаторов, рассматриваются стратегии взвешивания уровней значимости классификаторов [93], методы обучения правил комбинирования, учитывающие статистические особенности объединяемых классификаторов [94; 96] и методы, не привязанные к статистическим особенностям классификаторов, но использующие аппарат мультимножеств для построения модели групповой классификации объектов [97; 98].

Главным отличием видеопотока как цифрового образа распознаваемого объекта является тот факт, что для одного и того же объекта рассматривается последовательность наблюдений, которые отличаются между собой. Рассмотрим причины, по которым результат распознавания объекта может быть ошибочным, исходя из предположения, что система действует всегда детерми-

нировано, т.е. в любой момент времени и при любых внешних условиях результаты распознавания одного и того же набора входных данных всегда совпадают. Таким образом любая ошибка является следствием неспособности системы различить объект того или иного класса. Ошибки распознавания можно условно разделить на три группы:

1. *Ошибки, обусловленные несовершенством алгоритма распознавания,* т.е. ошибки, являющиеся «внутренними» с точки зрения системы распознавания объектов и которые могут проявляться даже при идеальном функционировании других подсистем. Данный класс ошибок является безусловным атрибутом любой системы распознавания, вне зависимости от модели входа.
2. *Ошибки, обусловленные дефектами предварительной обработки.* Система распознавания одиночного изображения, как правило, является одной из подсистем некоторого комплекса и изображения, подаваемые на вход системе распознавания формируются в результате действия других подсистем (см. рис. 2.3). Как следствие, могут возникнуть ошибки, связанные с несовершенством предшествующих подсистем. К примеру, пусть в результате разбиения изображения текстовой строки на изображения отдельных символов была допущена ошибка, в следствии которой положение правой границы изображения латинской буквы «P» было найдено некорректно, в результате чего на изображении буквы была утеряна перемычка между двумя горизонтальными штрихами. Изображение, полученное в результате, с точки зрения системы распознавания одиночного символа, может быть неотличимо от латинской буквы «F».
3. *Ошибки, обусловленные шумом среды.* Возникают такие ошибки в случае, если в условиях внешней среды, в которой находится распознаваемый объект, изображение этого объекта становится неотличимым от изображения объекта другого класса. К примеру, предположим, что производится съемка фотографии документа, удостоверяющего личность, содержащего поле «Имя» с истинным значением «HANNA». Данное поле начертано на белом фоне и документ покрыт защитной глянцевой поверхностью. В момент съемки на документе проявился блик от внешнего источника света, полностью закрывший букву «H» и оставивший изображения остальных букв неизменными. Таким образом, изоб-

ражение данного поля будет неотличимо от изображение поля «ANNA» на аналогичном документе.

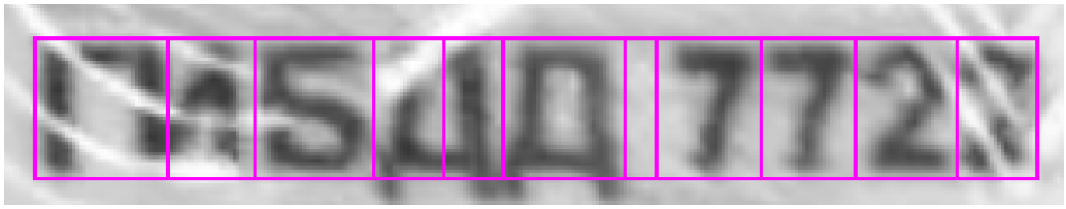


Рисунок 2.3 — Пример ошибочной сегментации текстовой строки на отдельные символы в условиях размытости изображения и дефектов, связанных с защитным голографическим слоем документа.

По отношению к системе распознавания одиночного изображения ошибки, связанные с шумом среды либо с дефектами предварительной обработки, являются следствием искажения входного изображения. Обладая возможность использовать несколько наблюдений объекта можно ожидать, что влияние шума среды и дефектов предварительной обработки на эти наблюдения будут различны. Однако даже при фиксировании системы распознавания одиночного объекта, вне зависимости от предварительной обработки, остаются ошибки, обусловленные несовершенством модели классификации. Современные исследования показывают, что наиболее высокоэффективный метод распознавания изображений, сверточные нейронные сети [63; 64], который в ряде отдельных задач показывает результаты, способные конкурировать с человеком [61], тем не менее может показывать неустойчивый результат при минимальных изменениях входного изображения [70; 71], даже если эти изменения касались всего лишь одного пикселя [72]. Так, даже используя наиболее точный метод распознавания, но обладая единственным входным изображением объекта, невозможно отделить полезный сигнал от шума, влияние которого может кардинальным образом поменять результат.

Таким образом, рассматривая в качестве цифрового образа объекта не одиночное изображение, а видеопоток, появляется возможность уменьшить влияние ошибок за счет вариативности шума применительно к отдельным кадрам видеопотока, которой не обладают классические системы распознавания объектов.

Одним из методов, позволяющих производить анализ множества изображений одной и той же сцены с целью уменьшить влияние шума оптической системы и дефектов, связанных с неконтролируемыми условиями съемки, яв-

ляется техника «супер-разрешения» — процесс получения изображения высокого разрешения из нескольких изображений того же объекта с более низким разрешением. Данной задаче уделялось большое внимание в литературе и предложено большое количество подходов, принимающих во внимание специфику финальной задачи обработки изображения и распознавания объекта или сцены [117; 118]. Однако как было отмечено ранее, дальнейшая обработка полученного единого изображения объекта остается подверженной ошибкам алгоритма распознавания, в частности, неустойчивости сверточных нейронных сетей.

2.2 Модель системы распознавания объектов в видеопотоке

Рассмотрим модель системы распознавания одиночного объекта x. Пусть задано множество, содержащее K классов $C = \{c_1, c_2, \ldots, c_K\}$. К примеру, рассматривая задачу распознавания отдельных символов поля «Фамилия» паспорта гражданина Российской Федерации, множество классов представляет собой русский алфавит с добавленными к нему символами пробела и дефиса. Рассматривая задачу типизации страницы документа на изображении после локализации ее границ и проективного исправления, множеством классом может выступать коллекция типов страниц документов, доступных для дальнейшей обработки. Отдельно следует упомянуть, что иногда в задачах распознавания объектов и явлений допускается наличие «пустого класса», который должен быть ответом системы распознавания на входное изображение объекта, о котором системе не известно, либо на изображение, которое не содержит объекта.

Пусть задано изображение объекта $I(x)$ из некоторого множества всевозможных изображений $\mathbb{I}$ и в рамках модели взаимодействия системы распознавания с пользователем/оператором (либо с другими компонентами системы) существует класс $c^*(x) \in C$, к которому принадлежит объект x. Задача распознавания изображения одиночного объекта состоит в определении этого класса. Результат работы системы распознавания в общем виде представим как всюду определенное отображение из множества классов C в множество оценок принадлежности: $\hat{f} : C \to \mathbb{R}$. Учитывая, что множество классов C содержит ровно K элементов:

$$\hat{f}\left(I(x)\right) = \left\{(c_1, q_1), (c_2, q_2), \ldots, (c_K, q_K)\right\}, \tag{2.1}$$

где $q_i \in \mathbb{R}, \quad i \in \{1, \ldots, K\}$ — вещественные оценки принадлежности объекта x к классу $c_i \in C$ при условии, что наблюдается изображение объекта $I(x)$. В качестве окончательного решения классификации принимается класс $c^*\left(I(x)\right) = \arg\max \hat{f}\left(I(x)\right)$. Тривиальная схема системы распознавания объекта в рамках описанной модели представлена на рис. 2.4.

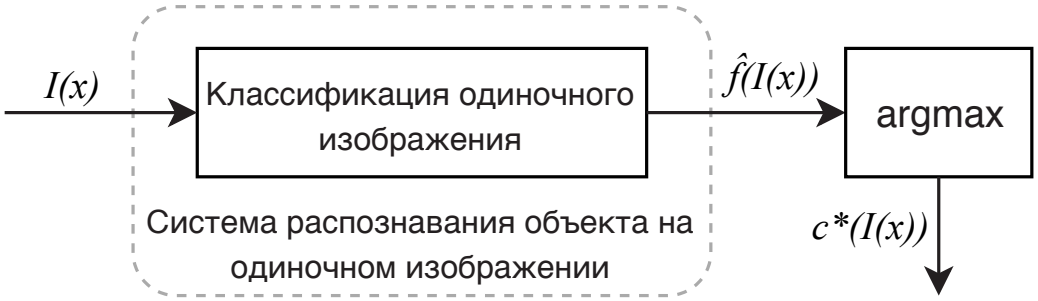


Рисунок 2.4 — Тривиальная схема системы распознавания одиночного объекта.

Если исключить из рассмотрения процесс валидации результатов распознавания и процесс обучения параметров системы распознавания (в случае, если для решения задачи классификации используются методы машинного обучения, к примеру, искусственные нейронные сети), и рассматривать непосредственно процесс распознавания, то такая система распознавания является статической и не предполагает обратных связей.

Рассмотрим теперь задачу распознавания объекта x в видеопотоке. Видеопоток генерируется при помощи некоторого захватывающего устройства, предоставляющего последовательность кадров, каждый из которых является независимым изображением объекта x. В условиях фиксированного количества кадров можно рассматривать задачу распознавания объекта в видеопотоке как статическую систему, аналогичную представленной на рис. 2.4, но с более сложной моделью входа. Тогда последовательность из N кадров можно рассматривать как множество изображений объекта x: $\mathbf{I}(x) = \{I_1(x), I_2(x), \ldots, I_N(x)\} \subset \mathbb{I}$. При этом модель выхода системы остается неизменной.

Реализации такой системы могут отличаться подходами к интеграции данных. Возможно тривиальное рассмотрение процесса классификации как «черного ящика», обрабатывающего сразу множество изображений (схема на рис. 2.5а). Другие варианты частично или полностью используют методы распознавания одиночных изображений объекта и осуществляют интеграцию либо на

уровне входных изображений (рис. 2.5б), либо на уровне результатов распознавания каждого отдельного изображения (рис. 2.5в).

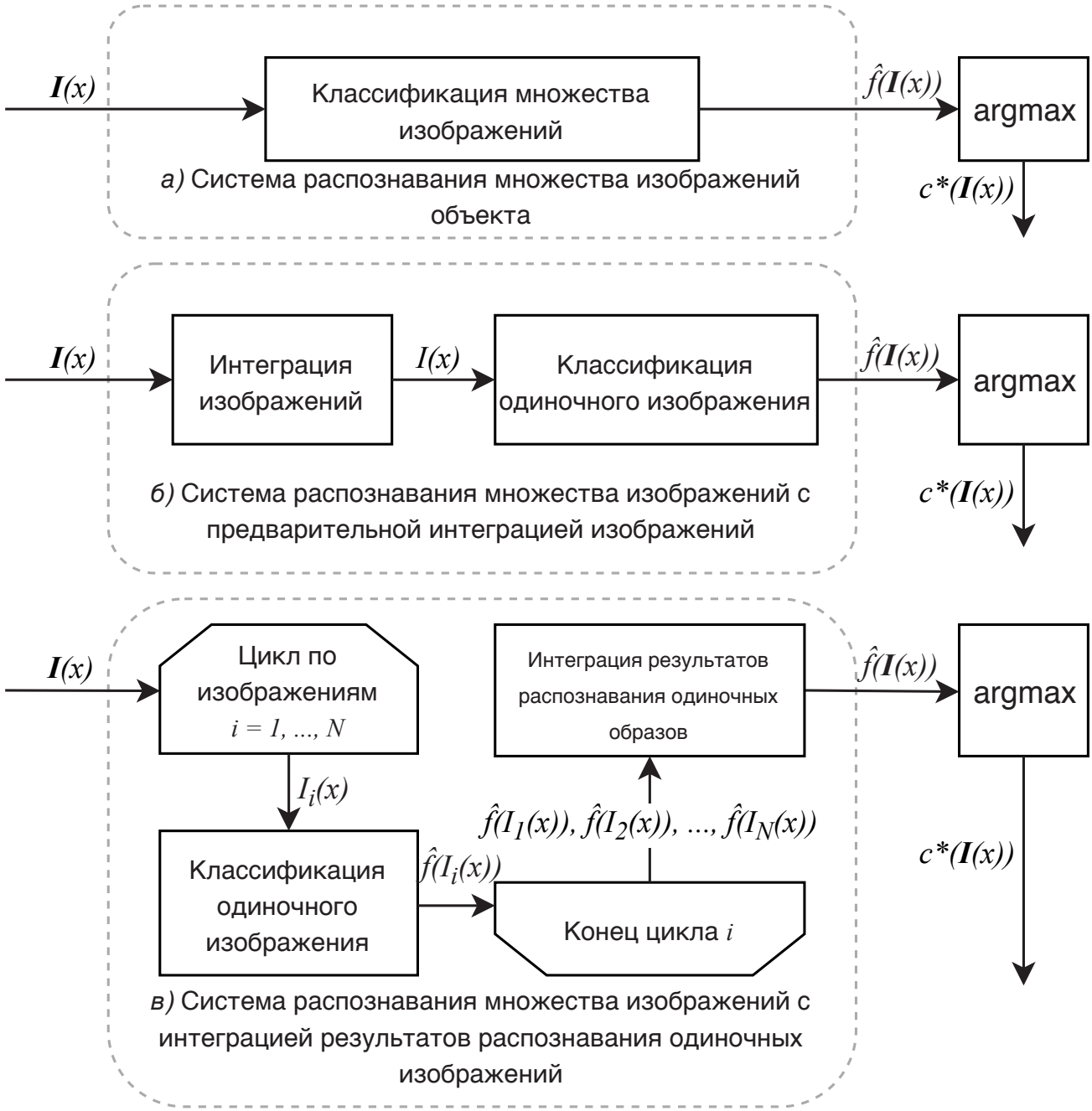


Рисунок 2.5 — Варианты статических систем распознавания множества изображений объекта.

Однако представленные статические модели системы распознавания объекта в видеопотоке не в полной мере отражают сценарий распознавания при помощи мобильного устройства — поскольку данные модели предполагают в качестве входа лишь множество кадров, без упорядочения, и не предполагают изменения состояния системы в процессе съемки. Также в условиях аппаратных ограничений мобильных устройств хранение и обработка множества изоб-

ражений может быть нецелесообразна или невозможна. Для того, чтобы более точно соответствовать процессу распознавания объекта в видеопотоке мобильного устройства предлагается рассмотреть динамическую модель с дискретным временем.

Для целей формализации представим видеопоток как генерирующаяся во времени последовательность изображений объекта. Таким образом, задано дискретное время $t = 0, 1, 2, \ldots$ и видеопоток, содержащий изображения наблюдаемого объекта $I_t(x) \in \mathbb{I}$. Подобная дискретная модель видеопотока соответствует принципам представления кодированного видеопотока в программных системах [119].

Для определения системы распознавания объекта в видеопотоке, который генерируется независимо, необходимо определить модель обслуживания, которая бы являлась промежуточным слоем между видеопотоком и непосредственным потоком обрабатываемых системой распознавания изображений. Наиболее тривиальной является схема обслуживания, при которой изображения, генерируемые во время обработки системой распознавания предыдущего изображения, сбрасываются. В случае, если возможно хранение коллекции изображений альтернативной моделью является схема обслуживания с буфером, позволяющим накапливать входящие изображения и выдавать их по запросу системы в произвольный момент времени, без ограничений, связанных с дискретизацией генерации изображений источником. С точки зрения непосредственно системы распознавания последовательности изображений набор методов и алгоритмов распознавания и интеграции результатов не зависят от схемы обслуживания, поэтому в рамках данной работы в дальнейшем будем предполагать, что в любой момент времени t может быть захвачено «текущее» изображение $I_t(x)$, а в периоды загрузки системы изображения могут сбрасываться.

Система распознавания поддерживает некоторое внутреннее состояние $s_t \in \mathbb{S}$, изменяющееся во времени. Время Δ_t, необходимое для получения обновленного результата после ввода очередного образа $I_t(x)$, в общем случае является функцией от изображения и внутреннего состояния системы: $\Delta_t = \Delta(I_t(x), s_t)$, которая может быть невычислима в момент времени t. Результат распознавания, учитывающий информацию, которая содержится в изображении, которое было захвачено в момент времени t, может быть доступен только в момент времени $T(t) = t + \Delta_t$.

В начальный момент времени $t = 0$ инициализировано внутреннее состояние системы s_0. Пусть в момент времени t происходит захват изображения $I_t(x)$, которое подается на модуль распознавания $\hat{f}$. Результат распознавания $\hat{f}(I_t(x))$ становится доступным в момент времени $t' \geqslant t$ и регистрируется в модуле памяти системы (т.е. становится частью состояния $s_{t'}$). После этого происходит комбинирование результатов распознавания изображений объекта, накопленных на текущий момент, и в момент времени $T(t) \geqslant t'$ происходит вывод результата распознавания $R_{T(t)}$. После вывода результата происходит захват очередного изображения $I_{T(t)}(x)$ и процесс продолжается. Таким образом, результат $R_{T(t)}$ учитывает информацию, которая содержится в изображениях с индексами $0, T^1(0), T^2(0), \ldots, t$ (под надстрочным знаком функции $T(t)$ подразумевается не возведение в степень, а множественная композиция функции). Качество результата характеризуется близостью результата $R_{T(t)}$ к истинному значению $\boldsymbol{\nu}(x)$ объекта x, согласно некоторой метрике. Схема описанной системы распознавания представлена на рис. 2.6.

Методы выделения признаков и классификации объектов, применимые в статических системах (см. рис. 2.5) также применимы и в динамической модели, однако динамическая модель системы распознавания объекта в видеопотоке обладает рядом специфических свойств. В первую очередь необходимо отметить усиленное влияние производительности алгоритмов распознавания одиночного изображения на выход системы. Действительно, уменьшение времени Δ_t, необходимого для распознавания одного изображения $I_t(x)$, позволяет обработать большее количество информации об объекте x за одно и то же абсолютное время (т.е. за одно и то же время с точки зрения пользователя/оператора). Помимо этого в рамках подобной системы возникают задачи, нетипичные для традиционных систем распознавания объектов на изображениях. Первой такой задачей является получение результата $R_{T(t)}$ – задача комбинирования (интеграции) результатов распознавания одного и того же объекта на разных изображениях в единый результат. Второй задачей является останов процесса распознавания – поскольку захват изображений может быть не ограничен естественным образом, в момент времени $T(t)$ возникает задача принятия решения о том, что процесс захвата следует прекратить и накопленный к текущему моменту результат принять за окончательный.

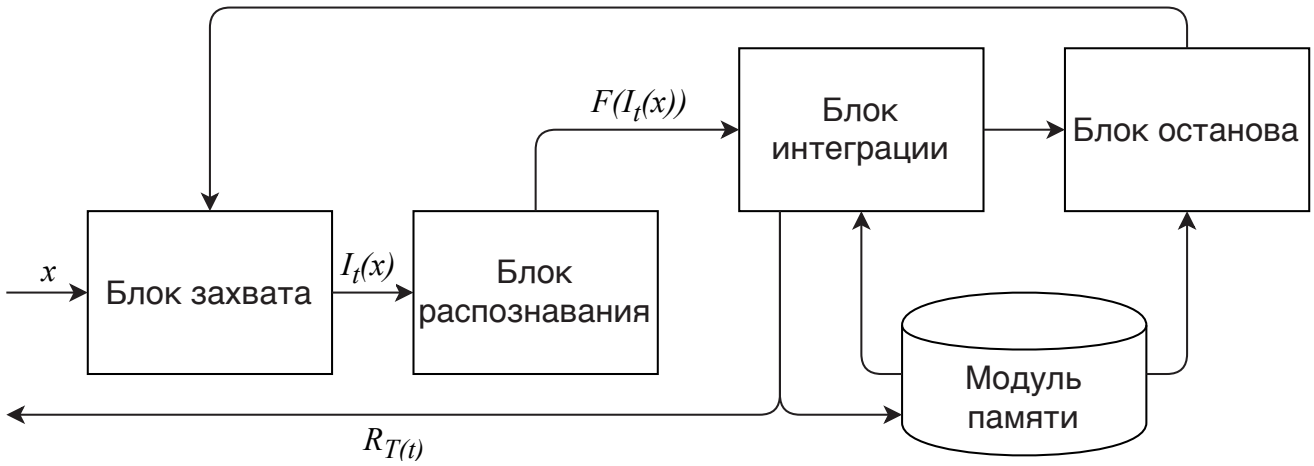


Рисунок 2.6 — Схема системы распознавания объекта в видеопотоке с остановом.

В качестве функционала эффективности системы в момент останова $t = t_{\text{stop}}$ предлагается рассматривать линейную комбинацию:

$$a \cdot \rho(R_{t_{\text{stop}}}, \nu(x)) + b \cdot W(t_{stop}), \tag{2.2}$$

где a, b – константы, $\rho(R_t, \nu(x))$ – расстояние от интегрированного результата R_t до истинного значения $\nu(x)$, характеризующая качество результата, а $W(t)$ – штрафная функция от времени. Частным случаем штрафной функции $W(t)$ является количество обработанных изображений:

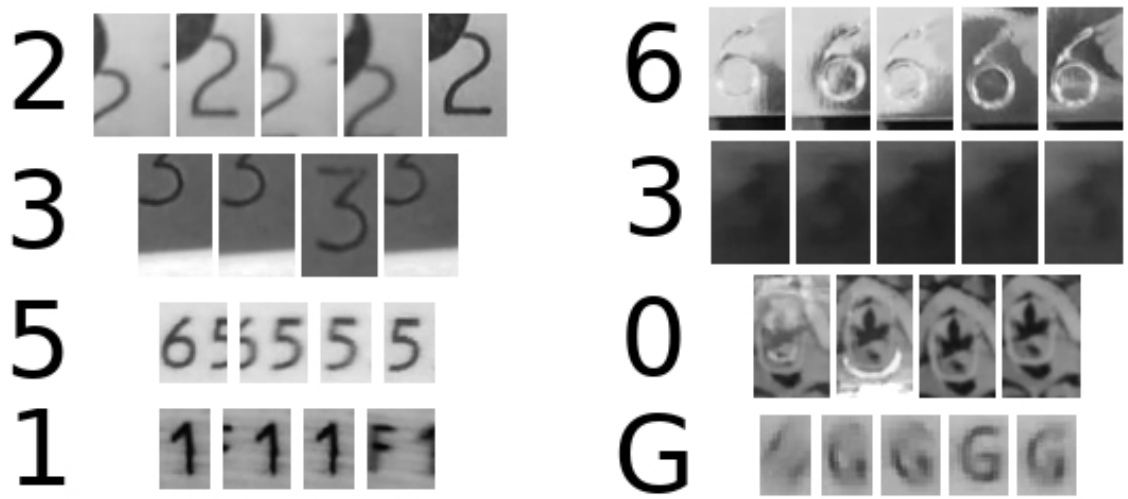
$$W(t) = \max\{i \mid T^i(0) \leqslant t\}. \tag{2.3}$$

2.3 Задача интеграции результатов распознавания объектов

Основной задачей традиционных систем распознавания объектов является максимизация точности распознавания (т.е. максимизация доли «правильных» классификаций объектов). *Задача интеграции результатов распознавания объектов* состоит в максимизации точности результата распознавания множества различных изображений одного и того же объекта при заданных результатах распознавания одиночных изображений.

На рисунке 2.7б представлены примеры последовательностей изображений одного и того же объекта, подверженные характерным искажениям, которые можно отнести к шуму среды: искажениям, связанным с оптической схемой

малоформатных цифровых камер, аберрациям, бликам и отражениям внутри оптической системы, цифровому шуму, неравномерной или недостаточной освещенностью сцены, расфокусировке изображения и «смазанности» ввиду движения оптического сенсора относительно носителя, бликам от внешнего источника освещения, геометрическим искажениям, таким как проективные искажения изображения объекта или нелинейные искажения, вызванные изгибами бумажного носителя, помехам, создаваемым голографическим защитным слоем и др. [15; 83] На рисунке 2.7а также представлены примеры последовательностей изображений объекта, подверженные дефектом предварительной обработки, в данном случае, ошибкам поиска и локализации объекта на входном кадре, ошибкам анализа структуры и локализации текстовых строк, ошибкам сегментации текстовых строк на отдельные символы [56].



а) б)

Рисунок 2.7 — Примеры последовательностей изображений объектов с дефектами предварительной обработки, порождающей изображение (а) и без дефектов предварительной обработки, но при воздействии шума среды (б).

Для формализации постановки задачи интеграции с точки зрения модели системы распознавания объекта в видеопотоке положим, что задан набор объектов $X = \{x_1, x_2, \ldots, x_M\}$ мощности M и набор видеопоследовательностей

$$B = \{\mathbf{I}_1(x_{b_1}), \mathbf{I}_2(x_{b_2}), \ldots, \mathbf{I}_H(x_{b_H})\} \tag{2.4}$$

мощности H, где b_h – индекс объекта из множества X для каждого $h \in \{1, 2, \ldots, H\}$, и каждая видеопоследовательность $\mathbf{I}_h(x_{b_h}) = \{I_{h1}(x_{b_h}), I_{h2}(x_{b_h}), \ldots, I_{hN_h}(x_{b_h})\}$ – последовательность изображений объекта

$x_{b_h} \in X$, которые могут быть подвержены шумам среды и дефектами предварительной обработки (см. раздел 2.1). Также задано множество классов $C = \{c_1, c_2, \ldots, c_K\}$ и информация об идеальной принадлежности каждого объекта к соответствующему классу $\nu : X \to C$.

Задачу распознавания объекта в видеопотоке можно сформулировать как поиск классифицирующей функции $F : \mathbb{I}^* \to C$, максимизирующей точность распознавания [120]:

$$V_F(B) = \frac{1}{H}\sum_{h=1}^{H}\left[F(\mathbf{I}_h(x_{b_h})) = \nu(x_{b_h})\right] \to \max_F . \tag{2.5}$$

Более частная задача интеграции результатов распознавания одиночных объектов предполагает функцию интегрирования результатов распознавания $R : (\mathbb{R}^C)^* \to \mathbb{R}^C$, преобразующую последовательность результатов распознавания одиночных изображений в единый результат распознавания видеопоследовательностей (здесь $\mathbb{R}^C$ — множество всевозможных отображений из множества классов C в множество оценок $\mathbb{R}$, т.е. множество всевозможных результатов классификации). Поскольку финальным ответом распознавания видеопоследовательности является класс, соответствующий максимальной оценке в результате распознавания $F(\mathbf{I}) = \arg\max R(\hat{f}(\mathbf{I}))$ (см. раздел 2.2), постановка задачи интеграции строится на основе (2.5) и приобретает вид:

$$V_R(B) = \frac{1}{H}\sum_{h=1}^{H}\left[\arg\max R\left(\hat{f}(\mathbf{I}_h(x_{b_h}))\right) = \nu(x_{b_h})\right] \to \max_R . \tag{2.6}$$

В идеальном случае классифицирующая функция F или функция интегрирования результатов R должна обладать возможностью фильтровать выбросы, появляющиеся во входном потоке данных из-за шума среды или дефектов предварительной обработки, и обладать возможностью проводить фильтрацию шума классификатора, нивелируя случайные внутренние ошибки.

Нетрудно заметить, что подход к интеграции как к задаче построения классифицирующей функции F можно свести к задаче построения функции R интегрирования результатов, применив имеющийся метод классификации одиночных изображений объектов. Альтернативными подходами являются, к примеру, техники «супер-разрешения» [118], осуществляющие пиксельное сопоставление множества входных изображений $\mathbf{I}(x)$ и построение единого «идеального» изображения объекта, которое впоследствии классифицируется. Однако

стоит заметить, что ввиду особенностей наиболее точного существующего в настоящий момент метода классификации изображений — сверточных нейронных сетей — а именно, его неустойчивости к случайным пиксельным искажениям, в рамках настоящей работы будут рассматриваться методы построения функции интеграции R результатов распознавания одиночных изображений.

В литературе задача объединения результатов классификации одиночных объектов обычно рассматривается в контексте методов получения более точной классификации путем объединения результатов нескольких разных классификаторов [89; 121; 122]. В зависимости от используемой модели результата классификации объекта и от интерпретации оценок классификатора рассматриваются различные методы комбинирования.

Задача комбинирования результатов классификации объектов можно быть рассмотрена как задача коллективного принятия решения. Введем понятие предиктора достоверности результата классификатора как вещественнозначную функцию $p(I(x), \hat{f})$, отражающую степень уверенности в том, что результат классификации изображения $I(x)$ функцией $\hat{f}$ будет верным. В качестве предикторов имеет смысл использовать вычислимые характеристики изображений, заведомо влияющие на точность классификации [123], такие как оценка смазывания и уровня фокусировки [115], оценка уровня шума, артефактов оцифровки [124] и пр. (такие предикторы можно считать *априорными*, поскольку они опираются непосредственно на характеристики входных изображений). Другой класс предикторов обуславливаются значениями оценок классификации (*апостериорные* предикторы), связанные с понятием оценки достоверности результата распознавания [7; 125]. Примером широко используемого апостериорного предиктора достоверности является значение оценки первой (максимальной) альтернативы [6]:

$$p(I(x), \hat{f}) = \max \hat{f}(I(x)). \tag{2.7}$$

Пусть задан некоторый предиктор достоверности. Тогда задача интеграции результатов распознавания последовательности изображений $\mathbf{I}(x) = \{I_1(x), \ldots, I_N(x)\}$ мощности N может быть рассмотрена как задача коллективного принятия решения с N экспертами, оценки уровней компетентности которых являются функциями от значений предиктора достоверности. Стоит заметить, что уровни компетентности экспертов в данной модели являются отражением входных данных – т.к. именно характеристики отдельных наблюде-

ний (т.е. отдельные изображения $I_1(x), \ldots, I_N(x)$) необходимы для оценки значимости экспертов.

Важным вопросом в рамках этой задачи является вопрос о целесообразности использования голосования нескольких экспертов вместо использования мнения самого компетентного эксперта [126; 127]. Переходя к частной задаче этот вопрос формулируется следующим образом: при каких моделях входных данных в задаче комбинирования результатов распознавания следует выбирать ту или иную стратегию комбинирования?

Для ответа на этот вопрос предлагается провести экспериментальное исследование. Были подготовлены четыре набора данных, характеристики которых приведены в таблице 2. Наборы данных MRZ-MSEGM и MRZ-CLEAN содержат видеопоследовательности результатов распознавания символов машиночитаемой зоны международных документов [83]. Наборы данных ICN-MSEGM и ICN-CLEAN содержат видеопоследовательности результатов распознавания символов поля «Номер» платежных банковских карт, выполненного при помощи индент-печати. Изображениям символов в рассматриваемых тестовых наборах свойственен широкий спектр искажений: неравномерная или недостаточная освещенность, цифровой шум, расфокусировка и «смазанность» ввиду движения оптического сенсора относительно носителя, блики от внешнего источника света и помехи, создаваемые голографическим защитным слоем документа и др. Результат распознавания каждого отдельного образа символа получен при помощи сверточных нейронных сетей, обученных отдельно для символов машиночитаемой зоны и для символов поля «Номер» платежных банковских карт, на отдельных обучающих наборах изображений с применением метода аугментации данных [128]. Наборы данных MRZ-MSEGM и ICN-MSEGM содержат ошибки, вызванные некорректной или недостаточно точной работой алгоритмов локализации документа и алгоритмов сегментации текстовых строк. Наборы MRZ-CLEAN и ICN-CLEAN являются подмножествами соответствующих наборов MRZ-MSEGM и ICN-MSEGM, не содержащим подобных ошибок. Таким образом, в наборах данных MRZ-CLEAN и ICN-CLEAN каждая видеопоследовательность содержит образы строго одного и того же символа, без каких-либо дефектов сегментации.

На представленных тестовых наборах данных проведено сравнение базовых стратегий комбинирования классификаторов, представленных в обзорной главе: правило произведения (1.2), суммы (1.3), минимума (1.4), максимума

Таблица 2 — Характеристики тестовых наборов данных MRZ-MSEGM, MRZ-CLEAN, ICN-MSEGM и ICN-CLEAN.

Характеристика набора данных	**MRZ-MSEGM**	**MRZ-CLEAN**
Мощность множества классов C	37	
Общее количество образов символов	637874	631530
Точность распознавания отдельных изображений, %	96.7357	96.8994
Количество видеопоследовательностей	7581	7508
Минимальная длина $\mathbf{I}(x)$	3	
Максимальная длина $\mathbf{I}(x)$	223	
Средняя длина $\mathbf{I}(x)$	21	
Характеристика набора данных	**ICN-MSEGM**	**ICN-CLEAN**
Мощность множества классов C	10	
Общее количество образов символов	31580	29166
Точность распознавания отдельных изображений, %	90.9816	96.8936
Количество видеопоследовательностей	1898	1748
Минимальная длина $\mathbf{I}(x)$	3	
Максимальная длина $\mathbf{I}(x)$	25	
Средняя длина $\mathbf{I}(x)$	12	

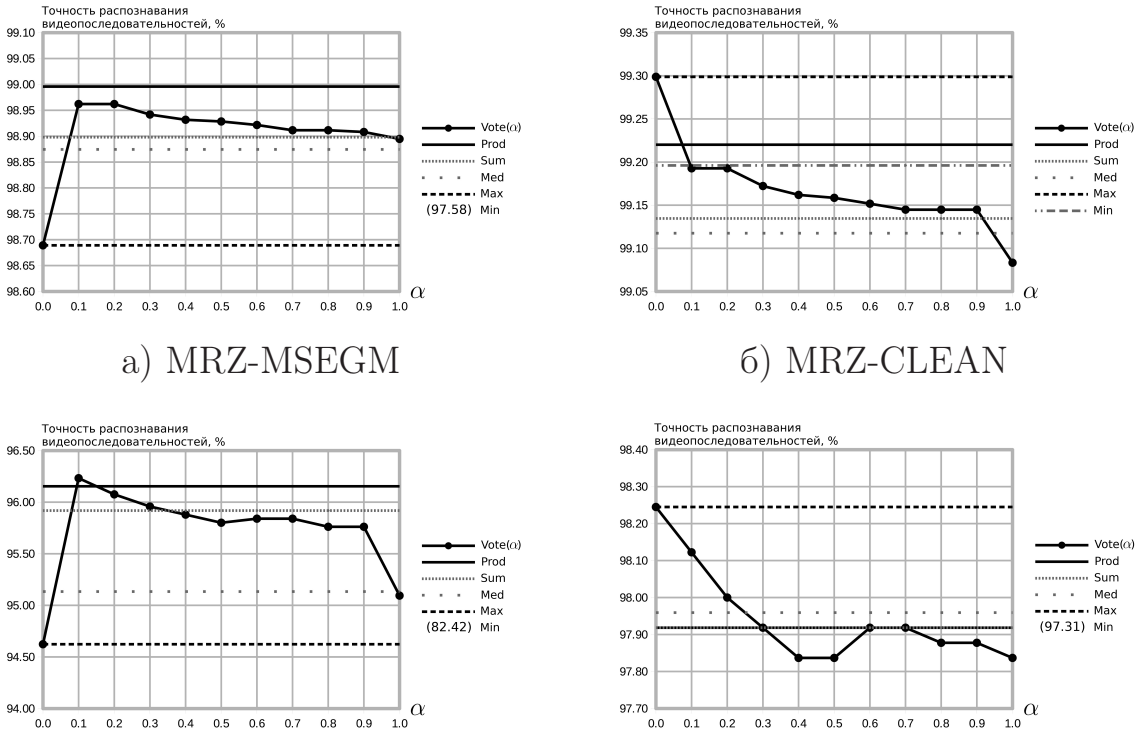
(1.5) и медианы (1.6). Точность распознавания видеопоследовательности символа является относительная доля видеопоследовательностей, для которых идеальный ответ совпадает с классом, получившим максимальную оценку согласно тому или иному правилу комбинирования. Дополнительно проведено сравнение базовых правил комбинирования с методом голосования (1.1), обобщенным следующим образом:

$$\begin{aligned}\mathrm{Vote}(\boldsymbol{\alpha})(\hat{f}(\mathbf{I}(x)))(c) = \\ = \boldsymbol{\alpha} \cdot \frac{1}{N}\sum_{i=1}^{N} 1_{\mathbf{I}_c(x)}(I_i(x)) + (1-\boldsymbol{\alpha}) \cdot \max_{i=1}^{N}\left(1_{\mathbf{I}_c(x)}(I_i(x)) \cdot p(I_i(x), \hat{f})\right), \quad (2.8)\end{aligned}$$

где $\mathbf{I}_c(x) = \{I(x) \in \mathbf{I}(x) \mid f(I(x)) = c\}$ – подмножество элементов видеопоследовательности, для которых выбором классификатора является класс c, $1_{\mathbf{I}_c(x)}(I(x))$ – индикаторная функция принадлежности образа $I(x)$ к подмножеству $\mathbf{I}_c(x)$, а $p(I(x), \hat{f})$ – предиктор достоверности. В качестве предиктора

достоверности использовался апостериорный предиктор «правило первой альтернативы» (2.7)

На рисунке 2.8 представлены сравнительные значения точности распознавания видеопоследовательностей с использованием правил комбинирования (1.2), (1.3), (1.4), (1.5), (1.6) и (1.1) на тестовых наборах данных MRZ-MSEGM, MRZ-CLEAN, ICN-MSEGM и ICN-CLEAN. Горизонтальная ось графиков соответствует значениям параметра α правила комбинирования (1.1). Точность распознавания с использованием остальных правил комбинирования представлены горизонтальными линиями.



а) MRZ-MSEGM б) MRZ-CLEAN

в) ICN-MSEGM г) ICN-CLEAN

Рисунок 2.8 — Сравнение точности распознавания видеопоследовательностей символов с использованием базовых стратегий комбинирования.

На рисунке 2.8 продемонстрирована значительная разница в оптимальном выборе стратегии комбинирования в зависимости от модели входных данных: на тестовых наборах, в которых встречаются ошибки локализации и сегментации символов, более высокую точность распознавания видеопоследовательностей обеспечивают правило произведения (1.2), голосование (1.1) и правило суммы (1.3) (рис. 2.8а, 2.8в). При этом на тестовых наборах, в которых такого типа ошибки были исключены (рис. 2.8б, 2.8г), более высокую точность распознавания обеспечивает правило максимума (1.5). Другими словами, при

рассмотрении данной задачи как задачи коллективного принятия решения, в случае более строгой модели входных данных (с отсутствием ошибок локализации и сегментации символов) выгоднее доверять единственному компетентному эксперту, нежели чем коллективному мнению нескольких экспертов.

При наличии ошибок локализации и сегментации символов устойчивость предикторов достоверности уменьшается, что в свою очередь увеличивает разницу между оценками компетентности экспертов (которые конструируются на основе значений предикторов) и действительными значениями компетентности экспертов (которые соответствуют апостериорным вероятностям принятия правильного решения). В таком случае выбор эксперта с максимальным уровнем компетентности чаще бывает ошибочным и таким образом разница между уровнем действительной компетентности выбранного эксперта и уровнями компетентности остальных экспертов сокращается. Таким образом оптимальность выбора наилучшего (с точки зрения устойчивого предиктора видеопоследовательности достоверности) отсутствуют покадрового ошибки результата локализации и в случае, сегментации когда в символов, соответствует более широкому положению теории коллективного принятия решения [126; 129], согласно которому нарушение первой части утверждения Кондорсе (при увеличении количества экспертов вероятность коллективного принятия правильного решения увеличивается, если для каждого эксперта вероятность принятия правильного индивидуального решения выше, чем вероятность принятия неправильного решения) происходит при увеличении разницы между уровнями компетентности максимально компетентного эксперта и остальных.

Из результатов проведенного эксперимента можно сделать вывод, что в случае построения системы распознавания объекта в видеопотоке для выбора стратегии комбинирования результатов необходимо руководствоваться не только моделью результатов распознавания объекта, но и моделью шума входных данных. При этом, в случае фиксированной модели шума входных данных для интеграции результатов классификации одиночных объектов можно пользоваться результатами исследований, которые были направлены на комбинирование различных классификаторов с целью максимизации точность распознавания одного объекта. В то же время, прямое применение рассмотренных правил комбинирования невозможно в случае, если модель результата распознавания объекта более сложна, чем простой результат классификации (2.1). В качестве примера такого объекта можно назвать текстовую строку, для которой клас-

сификация производится независимо для каждого символа. Задаче интеграции таких объектов будет посвящена глава 3.

2.4 Задача останова

Модель системы распознавания объекта в видеопотоке (см. рис. 2.6) не предполагает ограничения на количество входных изображений, а поскольку основной целью системы распознавания объектов является автоматизация ввода, важным параметром является абсолютное время (т.е. время с точки зрения оператора), необходимое для получения окончательного результата распознавания. В отличие от процесса съемки фотографии, видеопоток естественным образом не ограничен во времени. Отсюда следует *задача останова*, которая заключается в принятии решения о том, что вновь полученный результат $\hat{f}\left(\left\{I_0(x), I_{T^1(0)}(x), I_{T^2(0)}(x), \ldots, I_t(x)\right\}\right)$ в момент времени $T(t)$ можно считать окончательным и цикл захвата изображений можно прекратить. При распознавании сложных объектов, которые состоят из множества независимо распознаваемых объектов, решение об останове распознавания отдельных объектов влияет на время Δ_t, необходимое для распознавания составного объекта, а значит и на количество информации, обрабатываемой в рамках общей системы. Таким образом, задача останова (тесно связанная с задачей интеграции) является важным аспектом системы распознавания в видеопотоке, в особенности в рамках взаимодействия с другими подсистемами, объектом распознавания которых в совокупности является составной объект, такой как текстовое поле или документ в целом.

В простейшем виде правило останова можно представить в виде предиката, действующего на видеопоследовательности: $P : \mathbb{I}^* \to \{0, 1\}$. Истинность предиката влечет остановку процесса захвата и распознавания изображений:

$$P\left(\{I_1(x), I_2(x), \ldots, I_n(x)\}\right) = \begin{cases} 1 : \text{решение об останове;} \\ 0 : \text{продолжение работы.} \end{cases} \tag{2.9}$$

Пусть $\mathbf{I}(x) = \{I_1(x), I_2(x), \ldots, I_N(x)\}$ — последовательность изображений объекта $x \in X$, а $\mathbf{I}^{(n)}(x) = \{I_1(x), I_2(x), \ldots, I_n(x)\} \subseteq \mathbf{I}(x)$ — префикс этой последовательности, имеющий длину $n \leqslant N$. Обозначим через $D_P(\mathbf{I}(x))$ количество

изображений, которые будут обработаны системой распознавания до срабатывания правила останова (2.9):

$$D_P(\mathbf{I}(x)) = \min\left[N, \min\left\{|\mathbf{I}^{(n)}(x)| \;\middle|\; n \in \{1,2,\ldots,N\} \wedge P(\mathbf{I}^{(n)}(x))\right\}\right]. \quad (2.10)$$

С учетом правила останова при обработке видеопоследовательности $\mathbf{I}(x)$ на распознавание подаются только изображения из подпоследовательности $\mathbf{I}^{(P)}(x) = \mathbf{I}^{(D_P(\mathbf{I}(x)))}(x)$, и исходный набор видеопоследовательностей (2.4) принимает вид $B^{(P)} = \{\mathbf{I}_1^{(P)}(x_{b_1}), \mathbf{I}_2^{(P)}(x_{b_2}), \ldots, \mathbf{I}_H^{(P)}(x_{b_H})\}$.

Для формализации задачи останова воспользуемся общей моделью взаимодействия системы распознавания с пользователем, которая используется в задачах определения достоверности результата распознавания объекта [6; 81] и для оценки эффективности работы системы использует функционал, описанный в экономических терминах. Пусть W_c — стоимость ввода корректного результата распознавания объекта, W_e — стоимость ввода ошибочного результата, и W_f — стоимость распознавания одного изображения объекта. Тогда функция эффективности правила останова может быть записана в виде средней стоимости работы системы:

$$W_{F,P}(B) = W_c \cdot V_F(B^{(P)}) + W_e \cdot \left(1 - V_F(B^{(P)})\right) + W_f \cdot \frac{1}{H}\left(\sum_{h=1}^{H} D_P(\mathbf{I}(x))\right), \quad (2.11)$$

где $V_F(B^{(P)})$ — точность распознавания видеопоследовательностей с учетом останова по правилу P (2.9), вычисляемая согласно (2.5) (аналогично в случае интеграции результатов распознавания одиночных объектов точность вычисляется согласно (2.6)).

Упрощая выражение (2.11) и принимая во внимание константность W_e приходим к общей постановке задачи останова как к задаче поиска правила останова, оптимизирующего функционал эффективности:

$$W_{F,P}(B) = V_F(B^{(P)}) \cdot (W_c - W_e) + W_f \cdot \frac{1}{H}\left(\sum_{h=1}^{H} D_P(\mathbf{I}(x))\right) \to \min_P. \quad (2.12)$$

Аналогичный функционал эффективности строится с учетом функционала точности (2.6) в рамках задачи интеграции результатов распознавания одиночных объектов.

В контексте распознавания объектов в видеопотоке задача останова процесса распознавания является достаточно новой и малоизученной. Более подробно данная задача, а также метод ее решения, будут рассмотрены в главе 4.

2.5 Выводы по главе

В данной главе были показаны свойства задачи распознавания объекта в видеопотоке. Представлены различные способы формализации системы распознавания в видеопотоке и построена модель динамической системы с модулем комбинирования покадровых результатов распознавания и модулем останова. Были показаны свойства динамической системы распознавания объектов в видеопотоке и предложены постановки задачи интеграции результатов распознавания нескольких наблюдений одного и того же объекта и задачи останова в контексте таких систем.

Задача интеграции (комбинирования) результатов распознавания нескольких наблюдений одного и того же объекта рассмотрена как задача коллективного принятия решения. Показано, что модель входных данных может влиять на выбор оптимальной стратегии комбинирования. Согласно проведенному экспериментальному исследованию, на тестовых наборах данных, в которых встречаются дефекты предварительной обработки изображения, такие правила комбинирования как голосование и правило произведения. В то же время на тестовых наборах, в которых такого типа ошибки были исключены, более высокую точность распознавания видеопоследовательности обеспечивает правило максимума. В терминах коллективного принятия решения, в случае более строгой модели входных данных (т.е. при распознавании множества изображений с минимальным вкладом дефектов предварительной обработки) выгоднее доверять единственному наиболее компетентному эксперту, нежели чем коллективному мнению нескольких экспертов.

Описана задача останова процесса распознавания объекта в видеопотоке, возникающая ввиду отсутствия естественного ограничения на количество получаемых наблюдений во времени. Задача является новой применительно к системам оптического распознавания объектов.

Глава 3. Интеграция результатов распознавания строкового объекта в видеопотоке

3.1 Введение

Распознавание таких объектов как параграфы текста, текстовые строки, поля документов, и т.п., сопряжено с набором сложностей, в особенности если источником изображения является камера мобильного устройства. В подобных условиях съемки изображениям характерны искажения, такие как дефокусировка, смазывание, блики на светоотражающих поверхностях, недостаточное разрешения для достаточной точности алгоритмов распознавания символов и др. [15; 130; 131]. На рисунке 3.1 представлен пример блика на документе и его влияния на изображения текстовых полей, извлеченных из последовательных кадров видеопотока.

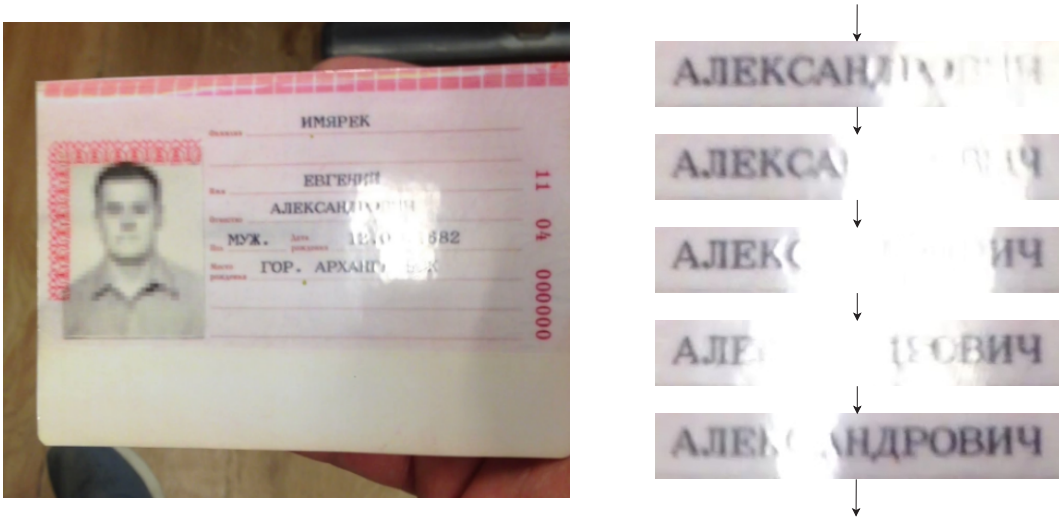


Рисунок 3.1 — Фрагмент кадра с бликом на отражающей поверхности документа (слева) и извлеченные изображения текстовых полей на кадрах видеопотока (справа). Изображения из пакета данных MIDV-500 [88] (клип HA39, поле 3)

Одним из преимуществ использования видеопотока при распознавании объектов является возможность обработки множества кадров в реальном времени, т.е. распознавания одного и того же объекта многократно, таким образом увеличивая финальную точность распознавания. Стоит также отметить,

что выбор единственного наилучшего результата в некоторых случаях может не являться приемлемой стратегией, так как в видеопотоке документа может не оказаться кадра с полностью видимым объектом. Таким образом, появляется необходимость в изучении методом комбинирования нескольких результатов распознавания.

Целями данной главы являются построение модели результата распознавания строкового объекта, учитывающей альтернативные варианты классификации одиночных объектов, и на ее основе построение алгоритма интеграции результатов распознавания строковых объектов. В разделе 3.2 будет описана модель результата распознавания одиночного и строкового объектов, используемая в дальнейшем для построения алгоритма. В разделе 3.3 приведена постановка задачи интеграции результатов распознавания строковых объектов. В разделе 3.4 приводится предлагаемый алгоритм, и в разделе 3.5 представлено его экспериментальное исследование.

3.2 Модель результата распознавания строкового объекта

Рассмотрим модель результата распознавания одиночного объекта. Пусть происходит классификация изображения I некоторого объекта c на один из K классов из множества $C = \{c_1, c_2, \ldots, c_K\}$ при помощи модуля классификации f. В классической постановке результатом классификации является один из классов $f(I) = c_f$, где $c_f \in C$, и задача распознавания одиночного объекта состоит в максимизации апостериорной вероятности совпадения класса c_f с истинным значением c. В более общей постановке модуль классификации $\hat{f}$ ставит входному изображению I в соответствие множество пар $\hat{f}(I) = \{(c_1, q_1), (c_2, q_2), \ldots, (c_K, q_K)\}$, где q_i – оценка принадлежности объекта к классу c_i. Финальным результатом распознавания является класс, соответствующий максимальной оценке принадлежности:

$$f(I) = \arg\max\{\hat{f}(I)\} \in \left\{ c_f \;\middle|\; \Big((c_f, q_f) \in \hat{f}(I)\Big) \wedge \left(q_f = \max_{(c,q)\in\hat{f}(I)} q \right) \right\}. \quad (3.1)$$

В случае, если существует несколько пар $(c_{f_1}, q_f), (c_{f_2}, q_f), \ldots$ с равным максимальным значением оценки принадлежности, в качестве ответа для берет-

ся один из классов согласно принятой конвенции (к примеру, класс с минимальным индексом в множестве C). Модель результата распознавания одиночного объекта (3.1) является вариантом модели результата Алгоритмов Вычисления Оценок (АВО, [132]) и также является наиболее широко используемой моделью в методах оптического распознавания изображений при помощи сверточных нейронных сетей [64].

Для определения результата распознавания строкового объекта необходимо ввести понятие пустого класса λ, обозначающего отсутствие одиночного объекта. Расширенным результатом классификации одиночного объекта будем считать отображение $a : C \cup \{\lambda\} \to [0, 1]$ из множества классов, объединенного с меткой пустого класса λ в множество оценок принадлежности. Каждая оценка принадлежности является вещественным числом от 0 до 1 и сумма оценок принадлежности равна единице. Таким образом задается множество всевозможных результатов распознавания одиночного объекта $\hat{C}$:

$$\hat{C} \stackrel{\text{def}}{=} \left\{ a \in [0,1]^{C \cup \{\lambda\}} \;\middle|\; \sum_{c \in C \cup \{\lambda\}} a(c) = 1 \right\}. \tag{3.2}$$

На множестве всевозможных результатов распознавания одиночного объекта $\hat{C}$ можно задать метрику следующим образом:

$$\rho_{\hat{C}}(a, b) \stackrel{\text{def}}{=} \frac{1}{2} \sum_{c \in C \cup \{\lambda\}} |a(c) - b(c)|, \quad \forall a, b \in \hat{C}. \tag{3.3}$$

Легко убедиться, что функция $\rho_{\hat{C}}(a, b)$ обладает свойствами действительной метрики:

1. $\rho_{\hat{C}}(a, b) = 0 \Leftrightarrow \forall c \in C \cup \{\lambda\} : a(c) = b(c) \Leftrightarrow a = b$, следовательно, аксиома тождества выполняется;
2. $\forall c \in C \cup \{\lambda\} : |a(c) - b(c)| = |b(c) - a(c)| \Rightarrow \rho_{\hat{C}}(a, b) = \rho_{\hat{C}}(b, a)$, следовательно, аксиома симметрии выполняется;
3. $\forall x, y \in \mathbb{R} : |x + y| \leqslant |x| + |y| \Rightarrow$
   $\Rightarrow \forall c \in C \cup \{\lambda\} : |a(c) - d(c)| \leqslant |a(c) - d(c)| + |d(c) - b(c)| \Rightarrow$
   $\Rightarrow \rho_{\hat{C}}(a, b) \leqslant \rho_{\hat{C}}(a, d) + \rho_{\hat{C}}(d, b)$, следовательно, неравенство треугольника также выполняется.

Стоит отметить, что метрика $\rho_{\hat{C}}(a, b)$ соответствует манхэттенской метрике в пространстве векторов над упорядоченным множеством $C \cup \{\lambda\}$. Поскольку

для a и b сумма значений при всех $c \in C \cup \{\lambda\}$ равна единице, множество значений функции $\rho_{\hat{C}}(a,b)$ является отрезком $[0,1]$.

Обозначим через $\hat{\lambda}$ «пустой результат»:

$$\hat{\lambda} \stackrel{\text{def}}{=} \{(\lambda,1),(c_1,0),(c_2,0),\ldots,(c_K,0)\}\,. \tag{3.4}$$

Результатом X распознавания строкового объекта будем называть строку над множеством $\hat{C} \setminus \{\hat{\lambda}\}$, т.е. элементом $X \in \mathbb{X}$, где $\mathbb{X} \stackrel{\text{def}}{=} (\hat{C} \setminus \{\hat{\lambda}\})^*$. Строка X представляет собой последовательность результатов распознавания одиночных объектов $X = x_1 x_2 \ldots x_n$, где $x_i \in \hat{C} \setminus \{\hat{\lambda}\}$, длиной строки $|X| = n$ называется количество элементов в этой последовательности. Обозначение $X_{i\ldots j}$ относится к подстроке строки X, включающей элементы $x_i x_{i+1} \ldots x_{j-1} x_j$ для $1 \leqslant i \leqslant j \leqslant n$. При $i > j$ подстрока $X_{i\ldots j}$ соответствует пустой строке $\hat{\lambda}$ нулевой длины.

Введем понятие элементарного редакционного изменения T как пары $(a,b) \neq (\hat{\lambda},\hat{\lambda})$, где $a,b \in \hat{C}$. Редакционное изменение $T = (a,b)$, применительно к строке X, соответствует:

1. замене элемента $x_i = a$ в строке X на элемент b, если $b \neq \hat{\lambda}$;
2. удалению элемента $x_i = a$ из строки X, если $b = \hat{\lambda}$;
3. вставке элемента b в строку X, если $a = \hat{\lambda}$.

Рассмотрим две произвольные строки $X, Y \in \mathbb{X}$ конечной длины. Редакционным предписанием называется последовательность элементарных редакционных изменений $T_{X,Y} = T_1 T_2 \ldots T_L$, переводящая строку X в строку Y. Весом редакционного предписания будем считать сумму расстояний (в терминах метрики $\rho_{\hat{C}}$) между парами объектов, участвующих в элементарных редакционных изменениях $T_i = (a_i, b_i)$ предписания $T_{X,Y}$:

$$w(T_{X,Y}) \stackrel{\text{def}}{=} \sum_{i=1}^{L} \rho_{\hat{C}}(a_i, b_i). \tag{3.5}$$

Метрика на множестве результатов распознавания строковых объектов $\mathbb{X}$ задается как минимальный вес редакционного предписания, переводящего одну строку в другую:

$$\rho_{\mathbb{X}}(X,Y) = \min\{w(T_{X,Y})\}. \tag{3.6}$$

Метрика $\rho_{\mathbb{X}}$ (3.6) может рассматриваться как одно из реализаций Обобщенного Расстояния Левенштейна (Generalized Levenshtein Distance, [133]), и обладает свойствами действительной метрики при условии, что $\rho_{\hat{C}}$ (3.3) также ими обладает [134].

Для расчета расстояния между двумя результатами распознавания строковых объектов $\rho_{\mathbb{X}}(X,Y)$ можно воспользоваться следующей рекуррентной схемой. Пусть $d(i,j) \stackrel{\text{def}}{=} \rho_{\mathbb{X}}(X_{1\ldots i}, Y_{1\ldots j})$ – расстояние между префиксами строк X и Y, имеющими длины i и j соответственно. Тогда

$$\begin{aligned} d(0,0) &= 0, \\ d(i,0) &= \sum_{k=1}^{i} \rho_{\hat{C}}(x_k, \hat{\lambda}), \\ d(0,j) &= \sum_{k=1}^{j} \rho_{\hat{C}}(\hat{\lambda}, y_k), \\ d(i,j) &= \min \left\{ \begin{array}{l} \rho_{\hat{C}}(x_i, \hat{\lambda}) + d(i-1,j), \\ \rho_{\hat{C}}(\hat{\lambda}, y_j) + d(i,j-1), \\ \rho_{\hat{C}}(x_i, y_j) + d(i-1,j-1) \end{array} \right\}, \end{aligned} \tag{3.7}$$

и искомому значению метрики $\rho_{\mathbb{X}}(X,Y)$ соответствует значение $d(|X|,|Y|)$.

Стоит отметить, что максимальным возможным значением метрики $\rho_{\mathbb{X}}(X,Y)$ является максимум длин строк X и Y (при использовании $\rho_{\hat{C}}$ (3.3) в качестве метрики на множестве результатов распознавания одиночных объектов). При этом, поскольку $\rho_{\mathbb{X}}$ является частным случаем Обобщенного Расстояния Левенштейна, существует способ построить нормализованный вариант этой метрики, с сохранением аксиом тождества, симметрии и неравенства треугольника [134]:

$$\tilde{\rho}_{\mathbb{X}}(X,Y) \stackrel{\text{def}}{=} \frac{2 \cdot \rho_{\mathbb{X}}(X,Y)}{\alpha \cdot (|X| + |Y|) + \rho_{\mathbb{X}}(X,Y)}, \tag{3.8}$$

где α – максимально возможный вес элементарной вставки или удаления. Для случая метрики $\rho_{\hat{C}}$ (3.3): $\alpha = \max\{\rho_{\hat{C}}(a,\hat{\lambda}), \rho_{\hat{C}}(\hat{\lambda},b), a,b \in \hat{C}\} = 1$.

Помимо Обобщенного Расстояния Левенштейна существуют и другие подходы к сравнению строковых объектов, такие как алгоритм динамической трансформации временной шкалы (Dynamic Time Warping, DTW [133; 135]). В классической постановке, однако, алгоритм DTW предполагает соответствие граничных компонентов строковых объектов, не предполагает штрафа за вставку и удаление компонентов, и не обладает свойствами метрики (не гарантирует выполнение неравенства треугольника).

3.3 Задача интеграции результатов распознавания строкового объекта

Рассмотрим задачу распознавания строкового объекта в видеопоследовательности. На вход системе подается последовательность изображений $I_1, I_2, \ldots, I_N$ строкового объекта $\boldsymbol{\nu} \in C^*$. При помощи модуля $\hat{F}$ распознавания строкового объекта на одиночном изображении каждому из изображений ставится в соответствие результат распознавания $\hat{F}(I_i) \in \mathbb{X}$. В рамках рассматриваемой модели будем полагать, что в исходном результате распознавания строкового объекта оценки принадлежности, соответствующие пустому классу $\boldsymbol{\lambda}$ равны нулю:

$$\begin{aligned} \hat{F}(I_i) &= X_i, \quad X_i \in \mathbb{X}, \\ X_i &= x_1^i x_2^i \ldots x_{n_i}^i, \\ x_j^i(\boldsymbol{\lambda}) &= 0, \quad \forall j \in \{1, \ldots, n_i\}. \end{aligned} \tag{3.9}$$

Задача состоит в комбинировании результатов $X_1, X_2, \ldots, X_N$ с некоторыми весами $w_1, w_2, \ldots, w_N$ в единый результат $X \in \mathbb{X}$, минимизирующий расстояние по некоторой метрике до истинного значения $\boldsymbol{\nu}$. Поскольку $X \in \mathbb{X}$ является строкой над множеством $\hat{C} \setminus \{\hat{\boldsymbol{\lambda}}\}$, а $\boldsymbol{\nu}$ – строкой над множеством классов C, для определения расстояния между ними необходимо провести дополнительную конвертацию. Наиболее естественным способом является приведение истинного значения $\boldsymbol{\nu}$ в вид строки $\hat{\boldsymbol{\nu}} \in \mathbb{X}$:

$$\begin{aligned} \boldsymbol{\nu} &= \boldsymbol{\nu}_1 \boldsymbol{\nu}_2 \ldots \boldsymbol{\nu}_{n_\nu}, \quad \boldsymbol{\nu}_j \in C \\ \hat{\boldsymbol{\nu}} &= \hat{\boldsymbol{\nu}}_1 \hat{\boldsymbol{\nu}}_2 \ldots \hat{\boldsymbol{\nu}}_{n_\nu}, \quad \hat{\boldsymbol{\nu}}_j \in \hat{C} \setminus \{\hat{\boldsymbol{\lambda}}\}, \\ \hat{\boldsymbol{\nu}}_j &\stackrel{\text{def}}{=} \{(\boldsymbol{\lambda}, 0), (c_1, 0), (c_2, 0), \ldots, (\boldsymbol{\nu}_j, 1), \ldots, (c_K, 0)\}, \end{aligned} \tag{3.10}$$

и в качестве расстояния от интегрированного результата X до истинного значения $\boldsymbol{\nu}$ использовать расстояние $\rho_{\mathbb{X}}(X, \hat{\boldsymbol{\nu}})$ (3.6), либо его нормализованный вариант $\tilde{\rho}_{\mathbb{X}}(X, \hat{\boldsymbol{\nu}})$ (3.8).

Однако, с точки зрения практического применения, важна также возможность получить финальный результат распознавания строкового объекта (по аналогии с финальным результатом (3.1) для одиночного объекта). Для получения финального результата можно воспользоваться следующей двухэтапной процедурой:

1. На первом этапе каждому компоненту $x_j \in \hat{C} \setminus \{\hat{\lambda}\}$ интегрированного результата $X = x_1 x_2 \dots x_{n_X}$ ставится в соответствие либо класс $c_{x_j} \in C$ с максимальной оценкой принадлежности $x_j(c_{x_j})$, либо пустой класс λ, если его оценка $x_j(\lambda)$ превышает некоторый порог θ:

$$\bar{x}_j = \begin{cases} \arg\max\limits_{c \in C} x_j(c), & \text{если } x_j(\lambda) < \theta, \\ \lambda, & \text{если } x_j(\lambda) \geqslant \theta. \end{cases} \tag{3.11}$$

2. На втором этапе из полученной строки $\bar{x}_1 \bar{x}_2 \dots \bar{x}_{n_X}$ удаляются все компоненты $\bar{x}_j = \lambda$. Результирующую строку $\bar{X}_\theta \in C^*$ можно использовать в качестве финального результата распознавания строкового объекта.

В качестве расстояния от интегрированного результата X до истинного значения ν теперь можно использовать расстояние Левенштейна $\text{levenshtein}(\bar{X}_\theta, \nu)$ [133] или его нормализованный вариант [134]:

$$\rho_L(\bar{X}_\theta, \nu) = \frac{2 \cdot \text{levenshtein}(\bar{X}_\theta, \nu)}{|\bar{X}_\theta| + |\nu| + \text{levenshtein}(\bar{X}_\theta, \nu)}. \tag{3.12}$$

Задача интеграции строковых объектов была рассмотрена в работе [87] в контексте распознавания речи. Вместо интеграции результатов распознавания нескольких изображений $I_1, I_2, \dots, I_N$ при помощи единого модуля распознавания $\hat{F}$, в [87] рассматривается интеграция результатов распознавания одного «изображения» I различными системами распознавания $F_1, F_2, \dots, F_N$. Данные постановки задач можно считать схожими с точностью до модели шума: интеграция результатов распознавания строкового объекта в видеопоследовательности направлена на фильтрацию компоненты шума в исходных изображениях $I_1, I_2, \dots, I_N$ (обусловленной неточностью входных данных, ошибками предварительной обработки и пр.) и ее влияние на результат работы модуля распознавания $\hat{F}$, тогда как интеграция результатов различных модулей распознавания направлена на фильтрацию шума, привнесенного самими модулями распознавания $F_1, F_2, \dots, F_N$.

Помимо распознавания речи, подход, представленный в [87], также применялся для комбинирования множества классификаторов в задачах оптического распознавания печатных [85] и рукописных [136] текстов.

Подход, описанный в [87] носит название ROVER (Recognizer Output Voting Error Reduction) и предполагает двухмодульную схему, представленную на рисунке 3.2. Не первом этапе *модуль выравнивания* приводит все входные

строковые объекты к виду строк одинаковой длины, производя соответствующие вставки пустого класса λ оптимальным образом. На втором этапе *модуль голосования* выбирает класс для каждого компонента результирующей строки на основе линейной комбинации частоты возникновения и оценки достоверности, порожденной модулем распознавания.

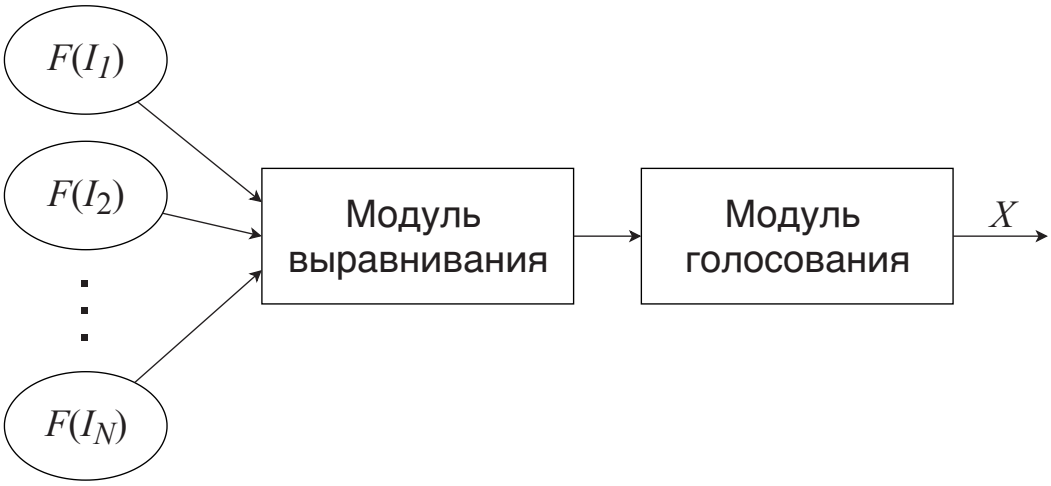


Рисунок 3.2 — Двухмодульная схема подхода ROVER [87]

Модель одиночного результата распознавания строки в подходе ROVER [87] представляет собой пару из строки над множеством классов распознавания одиночных объектов и оценки достоверности модуля распознавания, т.е. объект из множества $C^* \times \mathbb{R}$.

Для построения алгоритма интеграции результатов распознавания строкового объекта с расширенной моделью одиночного результата, рассмотрим постановку задачи выравнивания строк вида (3.9).

Пусть заданы N строк $X_1, \ldots, X_N$, где $X_i \in \mathbb{X}$, и $|X_i| = n_i > 0$:

$$\begin{aligned} X_1 &= x_1^1 x_2^1 \ldots x_{n_1}^1 \\ X_2 &= x_1^2 x_2^2 \ldots x_{n_2}^2 \\ &\ldots \\ X_N &= x_1^N x_2^N \ldots x_{n_N}^N \end{aligned} \tag{3.13}$$

Под выравниванием заданного множества строк будем понимать функцию $\text{align} : \{1, \ldots, N\} \times \{1, \ldots, \max\limits_{i=1}^{N} n_i\} \to \{1, \ldots, \sum_{i=1}^{N} n_i\}$. Функция $\text{align}(i, j)$ задает номер компонента выходной «интегрированной» строки, в значение которого вносит вклад компонент x_j^i. Для каждой входной строки значения функции align для отдельных компонент строки различны и сохраняют порядок: $\forall i \in \{1, \ldots, N\}, \forall j \in \{1, \ldots, n_i - 1\} : \text{align}(i, j) < \text{align}(i, j + 1)$.

Введем также функцию match : $\{1,\ldots,N\} \times \{1,\ldots,\sum_{i=1}^{N} n_i\} \to \hat{C}$, задаваемую следующим образом:

$$\mathrm{match}(i,k) \stackrel{\mathrm{def}}{=} \begin{cases} x_j^i, & \text{если } \mathrm{align}(i,j) = k, \\ \hat{\lambda}, & \text{если } \nexists j : \mathrm{align}(i,j) = k \end{cases}. \tag{3.14}$$

Задача выравнивания состоит в поиске функции выравнивания align такой, чтобы достичь минимального значения штрафного функционала:

$$\sum_{k} \sum_{i_1 < i_2} \rho_{\hat{C}}(\mathrm{match}(i_1,k), \mathrm{match}(i_2,k)) \to \min, \tag{3.15}$$

отражающего суммарное попарное расстояние между результатами распознавания одиночных объектов, вносящих вклад в одни и те же компоненты интегрированного результата.

Для обобщения модуля голосования (см. рис. 3.2), выбирающего класс для каждого компонента результирующей строки, введем семейство функций комбинирования результатов распознавания одиночных объектов $r^{(N)}$:

$$r^{(N)} : \hat{C}^N \times (\mathbb{R}_0^+)^N \to \hat{C} \setminus \{\hat{\lambda}\}. \tag{3.16}$$

Функция $r^{(N)}$ принимает на вход N результатов распознавания одиночных объектов $a_1, a_2, \ldots, a_N$ таких, что $\exists i : a_i \neq \hat{\lambda}$, и набор ассоциированных с ними неотрицательных весов $w_1, w_2, \ldots, w_N$, отражающих значимость результата, таких, что $\sum_{i=1}^{N} w_i > 0$.

Тогда функция интеграции результатов распознавания строковых объектов $R^{(N)}$ принимает вид:

$$R^{(N)}(X_1, X_2, \ldots, X_N, w_1, w_2, \ldots, w_N) = r_1^{(N)} r_2^{(N)} r_2^{(N)} \ldots r_{n_R}^{(N)}, \tag{3.17}$$

где $n_R = \max\limits_{i,j} \mathrm{align}(i,j)$, а каждая компонента результирующей строки вычисляется с использованием функции комбинирования (3.16) и в соответствии с результатом выравнивания (3.14):

$$r_j^{(N)} = r^{(N)}\left(\mathrm{match}(1,j)), \mathrm{match}(2,j), \ldots, \mathrm{match}(N,j), w_1, w_2, \ldots, w_N\right). \tag{3.18}$$

В общем случае точное решение задачи (3.15) предполагает расчет схемы динамического программирования (по аналогии со схемой расчета Обобщенного Расстояния Левенштейна (3.7)) с трудоемкостью, экспоненциально зависящей

от количества входных строк N (поскольку при расчете необходимо использовать результат подзадач выравнивания строк $X_{11\ldots i_1}, X_{21\ldots i_2}, \ldots, X_{N1\ldots i_N}$ для всех кортежей $(i_1, i_2, \ldots, i_N) \in \{1, \ldots, n_1\} \times \{1, \ldots, n_2\} \times \ldots \times \{1, \ldots, n_N\}$). При расчете данной схемы также можно использовать эвристические алгоритмы поиска кратчайшего пути, такие как A^*-поиск [137]. В следующем разделе будет представлен алгоритм интеграции результатов распознавания строковых объектов, с аппроксимацией функционала выравнивания методом, используемым в подходе ROVER [87].

3.4 Алгоритм интеграции результатов распознавания строкового объекта

При расчете интегрированного результата распознавания строкового объекта порождается набор промежуточных интегрированных результатов $R^{(1)}(X_1, w_1), \ldots, R^{(i-1)}(X_1, \ldots, X_{i-1}, w_1, \ldots, w_{i-1})$, где результат $R^{(i-1)}$ используется для решения задачи выравнивания на шаге i. На первом шаге алгоритма:

$$R^{(1)}(X_1, w_1) = X_1. \tag{3.19}$$

На каждом последующем i-м шаге алгоритма строится оптимальное выравнивание строк X_i и $R^{(i-1)}(X_1, \ldots, X_{i-1}, w_1, \ldots, w_{i-1})$ при помощи схемы динамического программирования, аналогичной (3.7). Пусть $d(l,m) \stackrel{\text{def}}{=} \rho_{\mathbb{X}}\left(X_{i1\ldots l}, R^{(i-1)}(X_1, \ldots, X_{i-1}, w_1, \ldots, w_{i-1})_{1\ldots m}\right)$, а $P_p(l, m)$ – вспомогательные функции для $p \in \{1,2,3\}$. Расчет $d(l,m)$ и $P_p(l, m)$ производится согласно следующей процедуре:

$$\begin{gathered} d(0,0) = 0, \quad d(l,0) = \sum_{k=1}^{l} \rho_{\hat{C}}(x_k^i, \hat{\lambda}), \quad d(0,m) = \sum_{k=1}^{m} \rho_{\hat{C}}(\hat{\lambda}, r_k^{(i-1)}), \\ \begin{aligned} P_1(l,m) &= \rho_{\hat{C}}(x_l^i, \hat{\lambda}) + d(l-1, m), \\ P_2(l,m) &= \rho_{\hat{C}}(\hat{\lambda}, r_m^{(i-1)}) + d(l, m-1), \\ P_3(l,m) &= \rho_{\hat{C}}(x_l^i, r_m^{(i-1)}) + d(l-1, m-1), \\ d(l,m) &= \min\{P_1(l,m), P_2(l,m), P_3(l,m)\}. \end{aligned} \end{gathered} \tag{3.20}$$

Для расчета результата интеграции на i-м шаге $R^{(i)}(X_1, \ldots, X_i, w_1, \ldots, w_i)$ введем две вспомогательные функции $t_X : \{0, \ldots, n_i + n_{R_{i-1}}\} \to \{1, \ldots, n_i\}$

и $t_R : \{0, \ldots, n_i + n_{R_{i-1}}\} \to \{1, \ldots, n_{R_{i-1}}\}$, расчет которых производится по следующей рекуррентной процедуре:

$$
\begin{aligned}
&t_X(0) = n_i,\\
&t_R(0) = n_{R_{i-1}},\\
&t_X(k+1) = \begin{cases} t_X(k), \text{ если } \begin{array}{c} P_2(t_X(k), t_R(k)) = d(t_X(k), t_R(k)) \wedge \\ \wedge\, P_1(t_X(k), t_R(k)) \neq d(t_X(k), t_R(k)) \end{array} \\ t_X(k) + 1, \text{ в остальных случаях,} \end{cases} \qquad (3.21)\\
&t_R(k+1) = \begin{cases} t_R(k), \text{ если } P_1(t_X(k), t_R(k)) = d(t_X(k), t_R(k)) \\ t_R(k) + 1, \text{ в остальных случаях.} \end{cases}
\end{aligned}
$$

Интегрированный результат на i-м шаге рассчитывается следующим образом:

$$
\begin{aligned}
&n_{R_i} = \min\{k : t_X(k) = t_R(k) = 0\},\\
&R^{(i)}(X_1, \ldots, X_i, w_1, \ldots, w_i) = r_1^{(i)} r_2^{(i)} \ldots r_{n_{R_i}}^{(i)},\\
&r_k^{(i)} = \begin{cases} r^{(2)}\left(r_{t_R(t(k))+1}^{(i-1)}, \hat{\lambda}, W_{i-1}, w_i\right), \text{ если } t_X(t(k)) = t_X(t(k)-1), \\ r^{(2)}\left(\hat{\lambda}, x_{t_X(t(k))+1}^{i}, W_{i-1}, w_i\right), \text{ если } t_R(t(k)) = t_R(t(k)-1), \\ r^{(2)}\left(r_{t_R(t(k))+1}^{(i-1)}, x_{t_X(t(k))+1}^{i}, W_{i-1}, w_i\right), \text{ в остальных случаях,} \end{cases}
\end{aligned} \qquad (3.22)
$$

где $W_i \stackrel{\text{def}}{=} \sum_{k=1}^{i} w_k$, вспомогательная функция $t(k) \stackrel{\text{def}}{=} n_{R_i} - k + 1$ а функция $r^{(2)}$ – функция интеграции двух результатов распознавания одиночных объектов (3.16).

Следует отметить, что рамках предлагаемого алгоритма от функции интеграции $r^{(N)}$ (3.16) требуется следующее свойство:

$$
\begin{aligned}
&r^{(N)}(a_1, \ldots, a_N, w_1, \ldots, w_N) = \\
&\quad = r^{(2)}(r^{(N-1)}(a_1, \ldots, a_{N-1}, w_1, \ldots, w_{N-1}), a_N, w_1 + \ldots + w_{N-1}, w_N). \qquad (3.23)
\end{aligned}
$$

В случае, если используемая функция r не обладает свойством (3.23), процедура выравнивания остается неизменной, а интегрированный результат на шаге i необходимо вычислять для каждого компонента результирующей строки по формуле (3.18), предварительно восстановив функции align и match (3.14) в явном виде.

В рамках данной диссертационной работы в качестве функции r предлагается использовать взвешенное среднее, обладающее свойством (3.23):

$$r^{(N)}(a_1, \ldots, a_N, w_1, \ldots, w_N)(c) = \frac{1}{W_N} \sum_{i=1}^{N} a_i(c) \cdot w_i, \quad \forall c \in C \cup \{\lambda\}. \tag{3.24}$$

В форме псевдокода процедура интеграции результатов распознавания строкового объекта представлена в виде Алгоритма 1.

Трудоемкость вычисления функций $\rho_{\hat{C}}$ (3.3) и r (3.24) составляет $O(K)$, где K – количество классов, на которое происходит классификация каждого одиночного объекта. Поскольку верхняя оценка на длину результирующей строки R после выполнения i-й итерации алгоритма составляет $O\left(\sum_{j=1}^{i} |X_i|\right) \leqslant O\left(i \cdot \max_{j=1}^{i} |X_i|\right)$, трудоемкость каждой итерации алгоритма можно оценить как $O(M^2NK)$, где $M = \max_{i=1}^{N} |X_i|$, и общую трудоемкость Алгоритма 1 как $O(M^2N^2K)$.

3.5 Экспериментальные результаты

В данном разделе будут продемонстрированы результаты экспериментального исследования работы алгоритма интеграции результатов распознавания строковых объектов, представленного в разделе 3.4. В качестве объекта распознавания рассматривалось текстовое поле документа, удостоверяющего личность.

Экспериментальное исследование проводилось на открытом пакете данных MIDV-500 [88], содержащем видеоролики 50 документов, удостоверяющих личность, различных типов (по 10 видеороликов для каждого документа, по 30 кадров в видеоролике) с размеченными идеальными позициями и значениями текстовых полей. Были проанализированы 4 группы полей: даты, записанные цифрами и знаками препинания, номер документа, строки машиночитаемой зоны (MRZ, Machine-Readable Zone) и компоненты имени держателя документы, записанные латинским алфавитом.

Рассматривались только кадры, на которых документ целиком присутствует в кадре (следовательно видеопоследовательности в рассматриваемом подмножестве пакета данных имели разную длину, от 1 до 30 кадров). Для

Алгоритм 1 Алгоритм интеграции результатов распознавания строкового объекта: расчет $R^{(N)}(X_1, X_2, \ldots, X_N, w_1, w_2, \ldots, w_N)$

Require: $N > 0$ **and** $\forall i \in \{1, \ldots, N\} : |X_i| > 0$

$R \leftarrow X_1$
$W \leftarrow w_1$
for $i = 2$ to N **do**
  $d(0,0) \leftarrow 0$
  $p(0,0) \leftarrow 0$ {метка пути}
  for $k = 1$ to $|X_i|$ **do**
    $d(k,0) \leftarrow d(k-1,0) + \rho_{\hat{C}}(x_k^i, \hat{\lambda})$ $\{X_i = x_1^i x_2^i \ldots x_{|X_i|}^i\}$
    $p(k,0) \leftarrow 1$ {путь 1 – выравнивание x_k^i и пустого символа}
  end for
  for $k = 1$ to $|R|$ **do**
    $d(0,k) \leftarrow d(0,k-1) + \rho_{\hat{C}}(\hat{\lambda}, r_k)$ $\{R = r_1 r_2 \ldots r_{|R|}\}$
    $p(k,0) \leftarrow 2$ {путь 2 – выравнивание r_k и пустого символа}
  end for
  for $l = 1$ to $|X_i|$ **do**
    for $m = 1$ to $|R|$ **do**
      $P_1 \leftarrow \rho_{\hat{C}}(x_l^i, \hat{\lambda}) + d(l-1, m)$
      $P_2 \leftarrow \rho_{\hat{C}}(\hat{\lambda}, r_m) + d(l, m-1)$
      $P_3 \leftarrow \rho_{\hat{C}}(x_l^i, r_m) + d(l-1, m-1)$
      $d(l,m) = \min\{P_1, P_2, P_3\}$
      if $P_1 = d(l,m)$ **then**
        $p(l,m) \leftarrow 1$
      else if $P_2 = d(l,m)$ **then**
        $p(l,m) \leftarrow 2$
      else
        $p(l,m) \leftarrow 3$ {путь 3 – выравнивание x_l^i и r_m}
      end if
    end for
  end for
  $R' \leftarrow \varnothing$ {пустая строка}
  $T_X \leftarrow |X_i|$
  $T_R \leftarrow |R|$
  while $T_X > 0$ **or** $T_R > 0$ **do**
    if $p(T_X, T_R) = 1$ **then**
      $R' \leftarrow r(\hat{\lambda}, x_{T_X}^i, W, w_i)R'$ {вставка нового элемента в начало R'}
      $T_X \leftarrow T_X - 1$
    else if $p(T_X, T_R) = 2$ **then**
      $R' \leftarrow r(r_{T_R}, \hat{\lambda}, W, w_i)R'$
      $T_R \leftarrow T_R - 1$
    else
      $R' \leftarrow r(r_{T_R}, x_{T_X}^i, W, w_i)R'$
      $T_X \leftarrow T_X - 1$
      $T_R \leftarrow T_R - 1$
    end if
  end while
  $R \leftarrow R'$
  $W \leftarrow W + w_i$
end for
return R

того, чтобы минимизировать эффекты нормализации и обеспечить более ясное представление результатов, каждый клип был дополнен до 30 кадров путем повторения клипа с начала (таким образом, все анализируемые клипы имели одну и ту же длину 30).

Каждое поле вырезалось из исходного изображения при помощи проективного преобразования, согласно совместной разметке идеальных границ документа и координат текстового поля, с добавленными отступами, равными 30% от наименьшей стороны текстового поля. Размер вырезаемых изображений текстовых полей соответствовал разрешению 300 точек на дюйм. Каждое вырезанное текстовое поле распознавалось при помощи компонента системы Smart IDReader [99], отвечающего за распознавание единичной текстовой строки, с расширенной моделью результата (3.9).

В качестве расстояния между интегрированным результатом распознавания текстового поля и его истинным значением использовалось нормализованное расстояние Левенштейна ρ_L (3.12) между истинным значением и текстовой строкой, полученной при помощи процедуры (3.11), описанной в разделе 3.3. Все сравнения значений символов проводились вне зависимости от регистра, а также латинская буква «O» считалась идентичной цифре «0».

В рамках данного экспериментального исследования Алгоритм 1, работающий в рамках расширенной модели результата распознавания строкового объекта, был сравнен с аналогом, работающим в рамках классической модели. Для каждой группы текстовых полей и для каждой видеопоследовательности проводилась интеграция методом ROVER, где в качестве входных данных использовались простые текстовые строки, сформированные процедурой (3.11), примененной к покадровым результатам распознавания. Порог θ значения оценки пустого символа (3.11) и для контрольного метода ROVER и для Алгоритма 1 был равен 0.6.

На рисунке 3.3 представлены результаты работы сравниваемых алгоритмов для четырех групп текстовых полей набора данных MIDV-500. Можно отметить, что для каждой группы полей интеграция Алгоритмом 1 полных результатов распознавания (т.е. с учетом альтернативных вариантов распознавания каждого одиночного символа) достигает меньшего значения ошибки чем интеграция методом ROVER (учитывающим только первые альтернативы распознавания каждого символа), вне зависимости от длины последовательности интегрируемых результатов.

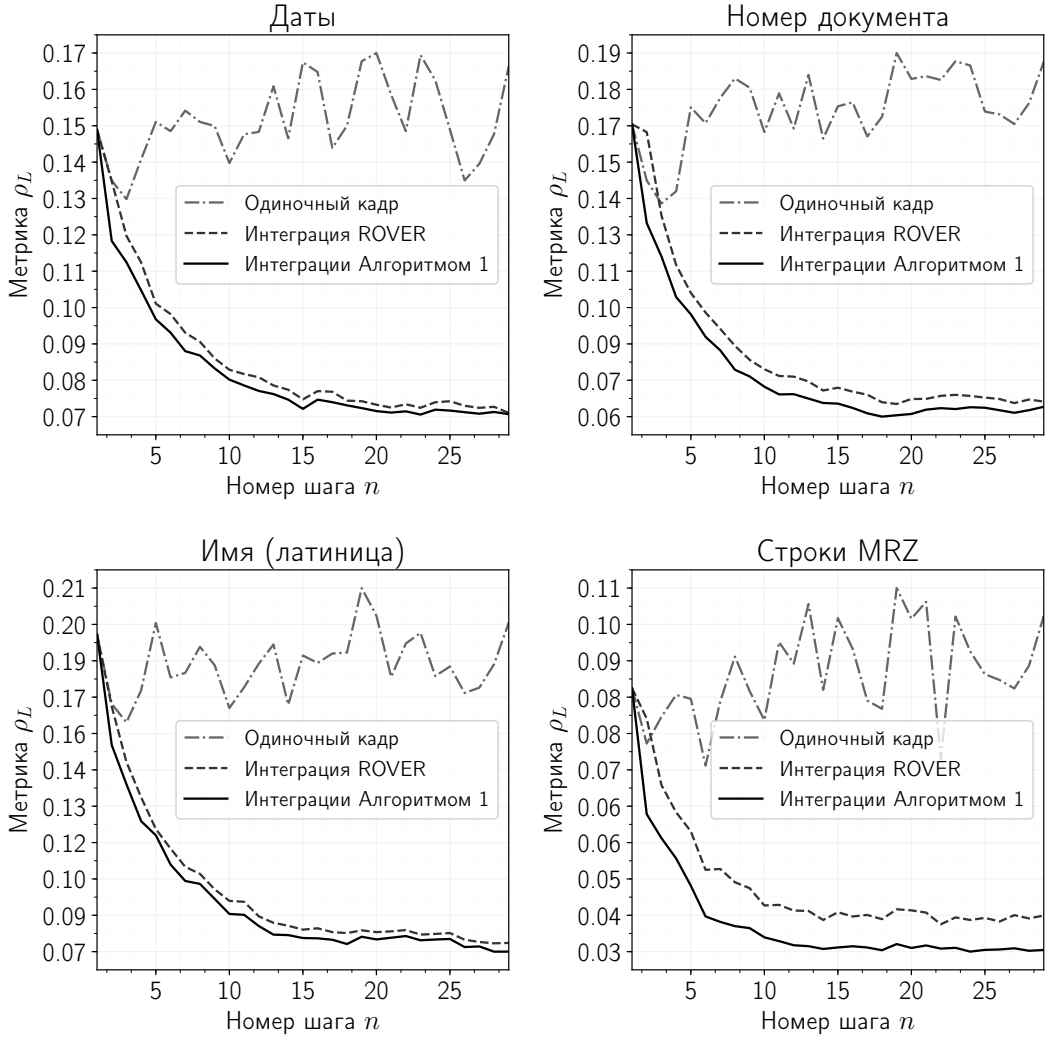


Рисунок 3.3 — Результаты работы алгоритмов интеграции для четырех групп текстовых полей набора данных MIDV-500

Таблица 3 — Достигнутое расстояние между интегрированным результатом распознавания и истинным значением без интеграции, методом ROVER и при помощи Алгоритма 1

Метод интеграции	Номер кадра (длина последовательности интегрируемых результатов)								
	3	6	9	12	15	18	21	24	27
Без интеграции	0.136	0.154	0.160	0.157	0.168	0.159	0.165	0.166	0.150
Интеграция методом ROVER	0.125	0.096	0.083	0.075	0.070	0.069	0.069	0.069	0.067
Интеграция Алгоритмом 1	**0.115**	**0.089**	**0.078**	**0.071**	**0.066**	**0.065**	**0.066**	**0.066**	**0.064**

На рисунке 3.4 представлены результаты работы алгоритмов совместно для всех четырех групп полей. Достигнутые средние значения расстояния между интегрированным результатом распознавания текстового поля и его истинным значениям для различных длин интегрируемого префикса видеопоследовательности представлены в таблице 3.

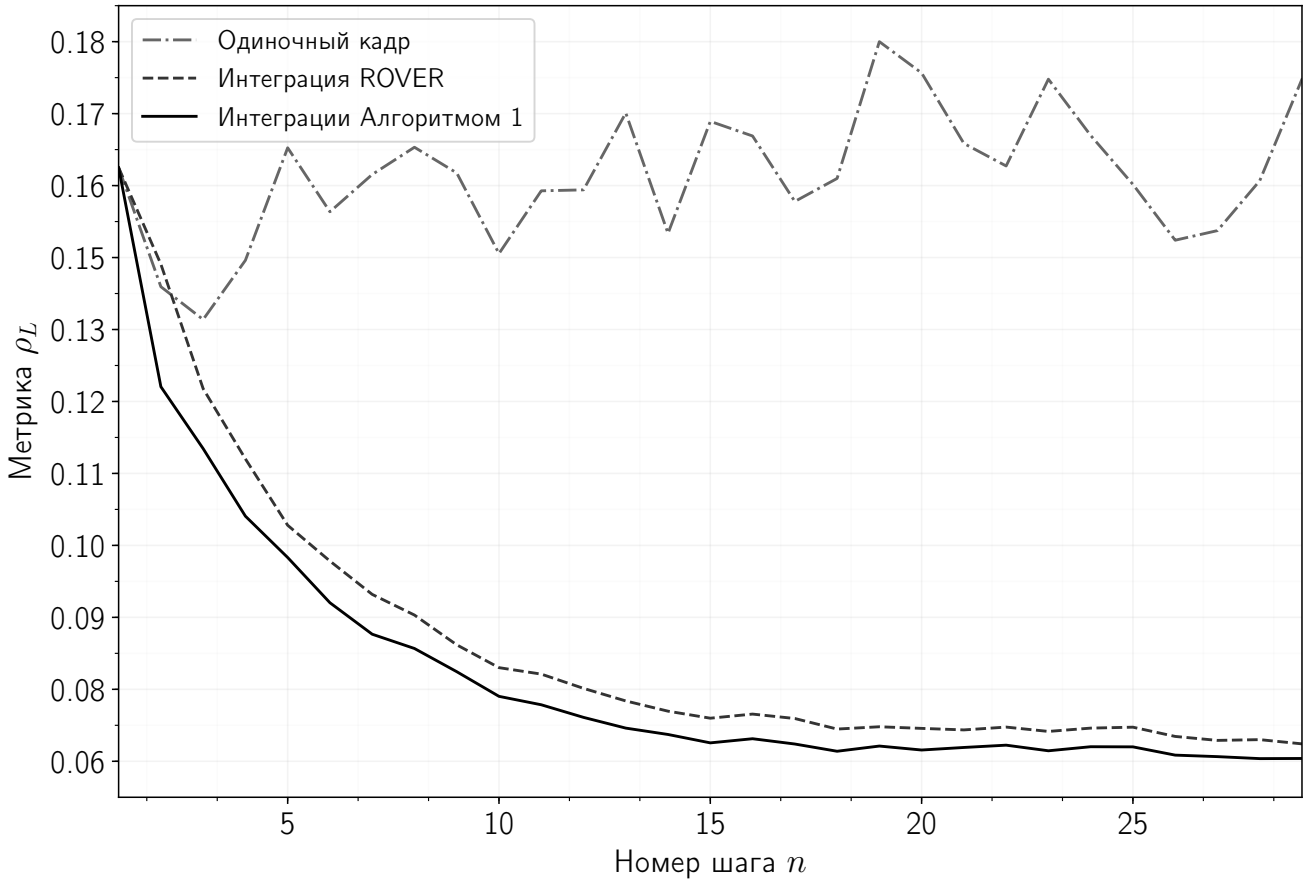


Рисунок 3.4 — Результаты работы алгоритмов интеграции для текстовых полей набора данных MIDV-500

3.6 Выводы по главе

В главе была рассмотрена задача комбинирования нескольких результатов распознавания строкового объекта с целью увеличения точности финального результата распознавания. Была описана модель результата распознавания строкового объекта, учитывающая альтернативные варианты классификации одиночных объектов и был представлен алгоритм интеграции результатов распознавания строковых объектов в рамках описанной модели.

По результатам проведенного экспериментального исследования можно сделать следующие выводы:

1. Методы интеграции результатов распознавания строковых объектов позволяют достичь значительного увеличения точности финального результата распознавания объекта при анализе видеопоследовательности.
2. Метод ROVER, в оригинале предназначающийся для комбинирования результатов распознавания одного и того же образа объекта несколькими алгоритмами распознавания, применим также для комбинирования результатов распознавания различных образов одного и того же объекта с использованием единого модуля распознавания.
3. И метод ROVER, принимающий на вход результаты распознавания строковых объектов в виде строк над множеством C классов значения одиночных объектов, так и Алгоритм 1, принимающий на вход результаты распознавания строковых объектов в расширенной модели (3.9), показывают значительное улучшение точности интегрированного результата при увеличении количества использованных кадров. В задаче распознавания текстовых полей документов, удостоверяющих личность, Алгоритм 1 показывает более высокую точность работы, чем прямое применение алгоритма ROVER.
4. По форме графиков зависимости расстояния между интегрированным результатом и истинным значением от количества использованных кадров (см. рис. 3.3 и 3.4) можно судить о том, что интеграция обладает свойством убывающей доходности (согласно терминологии алгоритмов «anytime» [102]). Это свойство является важным для решения задачи останова распознавания объекта в видеопоследовательности.

Глава 4. Задача останова процесса распознавания объекта в видеопотоке

4.1 Введение

Помимо задачи интеграции результатов распознавания объекта в условиях множественных наблюдений, при обработке видеопотока возникает задача оптимального останова. Проблема останова особенно актуальна применительно к системам компьютерного зрения, производящим распознавания объектов в реальном времени на мобильном устройстве [99]. В таких системах время получения результата распознавания зачастую так же важно, как и точность результата.

В данной главе будет рассмотрена задача останова процесса распознавания объекта в видеопотоке. В разделе 4.2 предлагается формальная постановка задачи останова распознавания объекта в видеопотоке как задачи принятия решения об останове итеративного процесса, в предположении, что оценки уверенности результатов распознавания недоступны. В разделе 4.3 описываются свойства монотонных задач останова, и в разделе 4.4 предлагается метод решения задачи в представленной постановке, путем ее сведения к монотонной задаче. В разделе 4.5 приводятся результаты экспериментального исследования, демонстрирующие эффективность предложенного метода для задачи распознавания текстового поля документа.

4.2 Формальная постановка задачи

Рассмотрим задачу распознавания объекта в видеопотоке. Пусть $\mathbb{X}$ обозначает множество всевозможных значений распознаваемого объекта (к примеру, множество строк над некоторым фиксированным алфавитом в случае распознавания текстовых полей), с заданной на нем метрикой $\rho : \mathbb{X} \times \mathbb{X} \to [0, +\infty)$. В видеопотоке производится распознавание объекта с истинным значением $X^* \in \mathbb{X}$. Процесс распознавания предполагает, что последовательность случайных ре-

зультатов распознавания $\mathbf{X} = (X_1, X_2, \ldots)$ наблюдается лицом, принимающим решение, один результат за один шаг процесса, и каждое наблюдение $x_i \in \mathbb{X}$ является реализацией X_i. Будем считать, что X_1,$X_2, \ldots$ имеют одинаковое совместное распределение с X^*. В рамках данной постановки мы предполагаем, что оценки достоверности результатов распознавания объекта недоступны.

Определим также семейство функций интеграции нескольких результатов распознавания, возвращающих в качестве значения единый интегрированный результат: $R^{(n)} : \mathbb{X}^n \to \mathbb{X}$. В любой момент n доступны результаты наблюдения $X_1 = x_1, \ldots, X_n = x_n$ и может быть получен интегрированный результат $R_n = R^{(n)}(x_1, \ldots,x_n)$. Процесс может быть остановлен в любое время $n > 0$ со следующей функцией штрафа:

$$c_e \cdot \rho(R_n, X^*) + c_f \cdot n, \tag{4.1}$$

где $c_e > 0$ обозначает стоимость ошибки распознавания в терминах расстояния до истинного значения, а $c_f > 0$ обозначает стоимость каждого наблюдения. Поскольку c_e и c_f являются положительными константами, без изменения структуры задачи оптимизации функция штрафа может быть определена следующим образом:

$$L_n \stackrel{\text{def}}{=} \rho(R_n, X^*) + c \cdot n, \quad c = c_f/c_e. \tag{4.2}$$

Значение функции штрафа при останове на шаге $n = 0$ (т.е. в случае, если ни одного наблюдения не было получено) можно считать бесконечным.

Задача состоит в выборе шага, на котором следует остановить процесс наблюдения так, чтобы минимизировать ожидаемый штраф. Формализуем эту постановку при помощи нотации, используемой в [106]. Правило останова может быть определено как последовательность функций:

$$\Phi \stackrel{\text{def}}{=} (\varphi_0, \varphi_1(x_1), \varphi_2(x_1, x_2), \varphi_3(x_1, x_2, x_3), \ldots)\,, \tag{4.3}$$

где $\forall n : 0 \leqslant \varphi_n(x_1, \ldots, x_n) \leqslant 1$. Функция $\varphi_n(x_1, \ldots,x_n)$ отражает условную вероятность останова на шаге n при условии, что шаг n был достигнут (т.е. при условии полученных наблюдений $X_1 = x_1, \ldots,X_n = x_n$).

Пользуясь правилом останова Φ и последовательность наблюдений $\mathbf{X}$ можно определить случайное время останова N. Обозначим через $P(N = n \mid \mathbf{X} = (x_1,x_2, \ldots))$ функцию вероятности останова на шаге n при заданной последовательности наблюдений $\mathbf{X}$. Эта функция выражается

через правило останова Φ (4.3) следующим образом:

$$\begin{aligned}
&P(N = 0 \mid \mathbf{X} = (x_1, x_2, \ldots)) = \varphi_0, \\
&P(N = n \mid \mathbf{X} = (x_1, x_2, \ldots)) = \varphi_n(x_1, \ldots, x_n) \times \\
&\qquad \times \prod_{j=1}^{n-1} (1 - \varphi_j(x_1, \ldots, x_j)) \quad \forall n \in \{1,2,\ldots\}, \\
&P(N = \infty \mid \mathbf{X} = (x_1, x_2, \ldots)) = 1 - \sum_{j=0}^{\infty} P(N = j \mid \mathbf{X} = (x_1, x_2, \ldots)).
\end{aligned} \tag{4.4}$$

В обратную сторону, при заданном случайном времени останова N правило останова для $n \in \{0,1,\ldots\}$ также может быть выражено в виде условной вероятности останова на шаге n при заданной последовательности наблюдений $\mathbf{X}$ и при условии, что процесс не останавливался на более ранних шагах:

$$\varphi_n(X_1, \ldots, X_n) = P(N = n | N \geqslant n, \mathbf{X} = (x_1, x_2, \ldots)). \tag{4.5}$$

Задача состоит в выборе правила останова Φ, доставляющего минимум функционалу ожидаемого убытка $V(\Phi)$, который может быть выражен следующим образом:

$$V(\Phi) = \mathrm{E}(L_N(X_1, \ldots, X_N)) \tag{4.6}$$

4.3 Оптимальный останов и монотонные задачи останова

4.3.1 Оптимальное правило останова

Поскольку метрика ρ не может принимать отрицательных значений, $\forall n : \; L_n \geqslant \; c \cdot n$, а также поскольку c является положительной константой, верны следующие утверждения:

$$\begin{aligned}
&\mathrm{E}(\inf_n L_n) > -\infty, \\
&\varliminf_{n \to \infty} L_n \geqslant L_\infty
\end{aligned} \tag{4.7}$$

В случае, если утверждения (4.7) верны можно показать [105; 106], что оптимальное правило останова существует и следует принципу оптимальности.

Обозначим через V_n^* минимальный ожидаемый убыток при любом правиле останова N таком, что $P(N \geqslant n) = 1$, т.е. при любом правиле останова, достигающем шага n:

$$V_n^*(x_1, \ldots, x_n) = \operatorname*{ess\,inf}_{N \geqslant n} \mathrm{E}_n(L_N), \tag{4.8}$$

где под $\mathrm{E}_n(\cdot)$ для краткости подразумевается условное ожидание при заданном наборе наблюдений вплоть до n-го шага $\mathrm{E}(\cdot \mid X_1 = x_1, \ldots, X_n = x_n)$. В (4.8) используется существенный инфимум, поскольку в общем случае существует более чем счетное множество правил останова $N \geqslant n$ и инфимум несчетного множества случайных величин может быть неизмерим [106]. Таким образом, (4.8) означает, что $P(V_n^*(x_1, \ldots, x_n) \leqslant \mathrm{E}_n(L_N)) = 1$ для всех $N \geqslant n$ и если Z – любая другая случайная величина, такая, что $\forall N \geqslant n : P(Z \leqslant \mathrm{E}_n(L_N)) = 1$, то $P(Z \leqslant V_n^*(x_1, \ldots, x_n)) = 1$.

Принцип оптимальности предполагает, что принимать решение об останове на шаге n оптимально тогда и только тогда, когда убыток в таком случае равен минимальному ожидаемому убытку для всех правил останова, достигающих шага n. Связь между минимальным ожидаемым убытком для всех правил останова $N \geqslant n$ и для всех правил останова, не останавливающихся на шаге n (т.е., достигающих шага $n + 1$) может быть выражена в виде равенства оптимальности:

$$V_n^* = \min\{L_n, \mathrm{E}_n(V_{n+1}^*)\}. \tag{4.9}$$

Пользуясь утверждениями (4.7) может быть доказано [106], что равенство (4.9) выполняется и что следующее правило останова является оптимальным:

$$N^* = \min\{n \geqslant 0 : L_n \leqslant \mathrm{E}_n(V_{n+1}^*)\}. \tag{4.10}$$

Другими словами, принцип оптимальности определяет правило (4.10), останавливающее процесс распознавания на первом шаге n, таком, что убыток при останове на нем не превышает ожидаемый убыток при любом правиле останова, который достигает шага n.

4.3.2 Монотонные задачи останова

Особый класс задач останова, класс *монотонных* задач [106; 116], определяется следующим образом. Пусть A_n обозначает событие $\{L_n \leqslant \mathrm{E}_n(L_{n+1})\}$. Задача останова называется монотонной, если выполняется:

$$A_0 \subset A_1 \subset A_2 \subset \ldots. \tag{4.11}$$

Условие (4.11) означает, что если на каком-то шаге n значение функции убытка не превосходит ожидаемый убыток на следующем шаге, то это же будет верно и на всех последующих шагах.

Рассмотрим следующее правило останова, которое носит название «близорукого правила» («myopic rule», «one-stage look-ahead rule»):

$$N_A = \min\{n \geqslant 0 : L_n \leqslant \mathrm{E}_n(L_{n+1})\}. \tag{4.12}$$

Правило N_A останавливает процесс на шаге n если текущее значение функции убытка не превосходит значения убытка при останове на шаге $n+1$. Можно показать [106; 116], что в случае, если задача останова монотонна и если у нее существует конечный горизонт (т.е. для некоторого фиксированного $T < \infty$ все правила останова должны останавливаться на шаге T), тогда близорукое правило (4.12) является оптимальным.

Для построения правила останова процесса распознавания объекта в видеопотоке в следующем разделе будут сформулированы условия, при которых задача можно считать монотонной, по крайней мере начиная с некоторого шага, и будет предложено правило останова, аппроксимирующее поведение близорукого правила (4.12).

4.4 Предлагаемый метод

Сформулируем следующее требование к функциям интеграции $R^{(n)}$: ожидаемое расстояние между двумя соседними интегрированными результатами распознавания не возрастает со временем:

$$\mathrm{E}(\rho(R_n,R_{n+1})) \geqslant \mathrm{E}(\rho(R_{n+1}, R_{n+2})) \quad \forall n > 0. \tag{4.13}$$

В терминологии алгоритмов «anytime» [102] требование (4.13) означает, что задача обладает свойством *убывающей доходности*. Пользуясь таким предположением о функциях интеграции $R^{(n)}$ можно показать, что задача останова (4.6) с функцией убытка (4.2) становится монотонной начиная с некоторого шага.

Действительно, обозначим через B_n событие $\{\mathrm{E}_n(\rho(R_n, R_{n+1})) \leqslant c\}$ и рассмотрим задачу останова (4.6) начиная с шага n, на котором событие B_n впервые произошло. События, рассматриваемые в условии монотонности (4.11) принимают следующий вид:

$$\begin{aligned} A_n : \{\rho(R_n, X^*) + cn \leqslant \mathrm{E}_n(\rho(R_{n+1}, X^*)) + cn + c\} = \\ = \{\rho(R_n, X^*) - \mathrm{E}_n(\rho(R_{n+1}, X^*)) \leqslant c\}. \end{aligned} \tag{4.14}$$

При фиксированном X^*, на шаге n, пользуясь неравенством треугольника можно получить соотношение между расстоянием от текущего результата распознавания до истинного значения, ожидаемым расстоянием до результата на следующем шаге и ожидаемым расстоянием от следующего результата до истинного значения:

$$\begin{aligned} \rho(R_n, X^*) \leqslant \mathrm{E}_n(\rho(R_n, R_{n+1})) + \mathrm{E}_n(\rho(R_{n+1}, X^*)) \Rightarrow \\ \Rightarrow \rho(R_n, X^*) - \mathrm{E}_n(\rho(R_{n+1}, X^*)) \leqslant \mathrm{E}_n(\rho(R_n, R_{n+1})). \end{aligned} \tag{4.15}$$

Если правая часть неравенства, полученного в (4.15) не превышает константы c, то и левая часть также не превышает c, и, следовательно, если происходит событие B_n, то и событие A_n (4.14) также должно произойти. Более того, используя предположение (4.13) мы можем получить, что если событие B_n произойдет, то и событие B_{n+1} также произойдет. Таким образом,

$$\forall n > 0 : \quad B_n \subset A_n, \quad B_n \subset B_{n+1}. \tag{4.16}$$

Из этого следует, что начиная с шага n, на котором событие B_n произошло впервые, события $A_n, A_{n+1}, A_{n+2} \ldots$ также произойдут, а значит задача останова может рассматриваться как монотонная начиная с этого шага, из чего следует оптимальность близорукого правила (4.12) среди всех правил останова, достигающих шага n в случае, если задача имеет конечный горизонт.

Рассмотрим теперь правило останова, предписывающее лицу, принимающему решение, остановить процесс распознавания в случае, если произошло событие B_n:

$$N_B = \min\{n > 0 : \mathrm{E}_n(\rho(R_n, R_{n+1})) \leqslant c\}. \tag{4.17}$$

Если правило N_B требует останова на шаге n, то и правило N_A потребует останова на этом шаге, а поскольку задача становится монотонной начиная с шага n, решение правила N_A является оптимальным, а значит и оптимальное правило N^* также потребует останова на этом шаге. Более того, если $\rho(R_n, X^*) - \mathrm{E}_n(\rho(R_{n+1}, X^*)) > c$, то правило N_B не останавливает процесс, также как и правило N^*, следующее принципу оптимальности. Следовательно, в случае если предположение (4.13) верно, правило N_B никогда не остановится раньше времени, и если правило требует останова, то решение об останове оптимально.

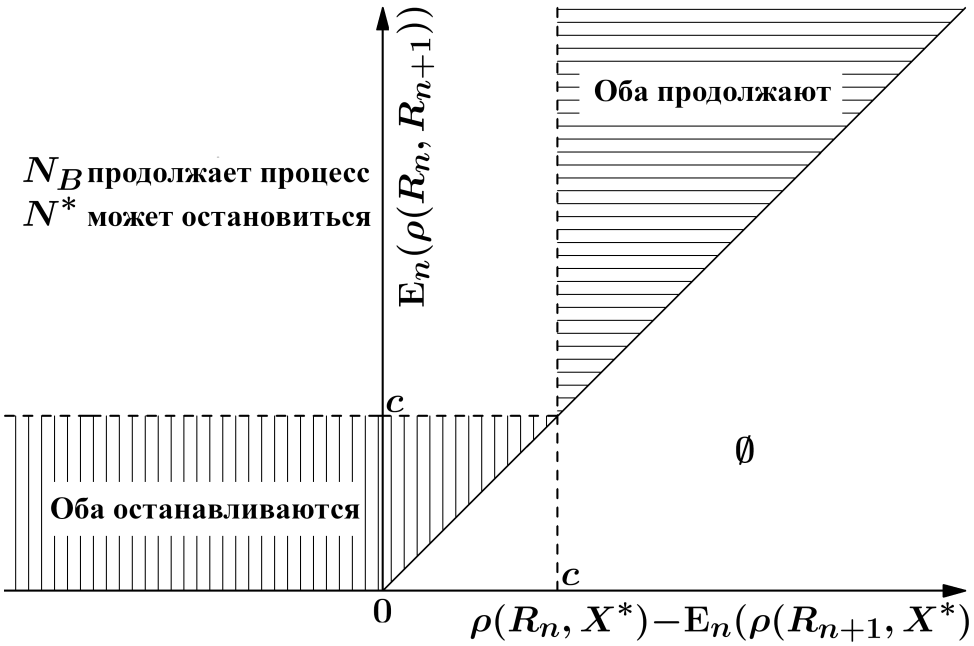


Рисунок 4.1 — Разница поведений предлагаемого правила останова N_B (основанного на оценке ожидаемого расстояния от текущего интегрированного результата распознавания до следующего) и оптимального правила останова N^*

На рисунке 4.1 графически показаны сходства и различия правил останова N_B и N^* при различных соотношениях между событиями A_n и B_n. Множество ситуаций, при которых N_B продолжает процесс, а N^* может остановить процесс, обусловлено двумя основными недостатками N_B:

1. Правило опирается на оценку разницы значений функции убытка используя неравенство треугольника, и, следовательно, является неэффективным в случае, если точность интегрированных результатов распознавания ухудшается со временем (т.е. если $\rho(R_n, X^*) - \mathrm{E}_n(\rho(R_{n+1}, X^*)) < 0$);

2. Решение принимается путем порогового отсечения ожидаемого значения метрики, что в общем случае может не быть ограничено сверху.

В предлагаемом методе будем предполагать, что метрика, задаваемая на множестве всевозможных результатов распознавания объекта, ограничена сверху (т.е. $\exists G : \forall x,y \in \mathbb{X} : 0 \leqslant \rho(x,y) \leqslant G$) и что функции интеграции $R^{(n)}$ порождают результаты, которые в среднем не ухудшаются во времени:

$$\mathrm{E}(\rho(R_n,X^*)) \geqslant \mathrm{E}(\rho(R_{n+1},X^*)) \quad \forall n > 0. \tag{4.18}$$

Тем самым, для решения задачи останова (4.6) с функционалом убытка (4.2) предлагается использовать следующий метод:

1. Оценить ожидаемое расстояние (в терминах метрики ρ) от текущего интегрированного результата распознавания объекта R_n (известного на шаге n) до неизвестного результата R_{n+1} на следующем шаге;
2. Принимать решение об останове процесса на шаге n, производя пороговое отсечение расстояния, оцененного в пункте 1, таким образом аппроксимируя поведение правила N_B.

В общем случае выбор метода прогнозирования следующего интегрированного результата распознавания объекта (или оценки ожидаемого расстояния между ним и текущим интегрированным результатом) может зависеть от природы функций интеграции $R^{(n)}$ и от других специфических характеристик задачи.

Построим на основе предложенного метода алгоритм останова процесса распознавания строкового объекта. Пусть заданы функции интеграции результатов распознавания строкового объекта $R^{(n)}$ (которые могут быть реализованы при помощи метода ROVER или при помощи Алгоритма 1). В качестве метрики ρ на строковых объектах предлагается использовать нормализованное обобщенное расстояние Левенштейна [134]. Для того, чтобы аппроксимировать поведение правила останова N_B, на n-м шаге процесса необходимо вычислять оценку ожидаемого расстояния между соседними интегрированными результатами распознавания $\Delta_n \stackrel{\text{def}}{=} \mathrm{E}_n(\rho(R_n, R_{n+1}))$, имея доступ к наблюдениям $X_1 = x_1, \ldots, X_n = x_n$. Для вычисления оценки предлагается провести моделирование следующего интегрированного результата исходя из предположения, что новое наблюдение будет близко к уже полученным на предыдущих шагах

наблюдениям:

$$\hat{\Delta}_n \stackrel{\text{def}}{=} \frac{1}{n+1}\left(\delta + \sum_{i=1}^{n} \rho(R_n, R(x_1,x_2,\ldots,x_n,x_i))\right), \tag{4.19}$$

где δ – внешний настраиваемый параметр.

В форме псевдокода алгоритм останова процесса распознавания строкового объекта представлен как Алгоритм 2. При использовании модели результата распознавания строкового объекта с альтернативными вариантами классификации одиночных объектов, рассмотренной в третьей главе, верхняя оценка длин интегрированных результатов R_n и R_{n+1} составляет $O(Mn)$, где $M = \max_{i=1}^{n} |X_i|$. Поскольку трудоемкость прямого вычисления обобщенного расстояния Левенштейна между строками X и Y составляет $O(|X| \cdot |Y| \cdot K)$, где K – количество классов, на которое происходит классификация каждого одиночного объекта, трудоемкость Алгоритма 2 составляет $O(M^2 n^3 K)$. Следует отметить, что трудоемкость алгоритма может быть снижена как путем использования упрощенных моделей результата распознавания строкового объекта, так и при помощи эвристических алгоритмов приближенного вычисления обобщенного расстояния Левенштейна.

Алгоритм 2 Алгоритм принятия решения об останове процесса распознавания строкового объекта на основе полученных наблюдений $X_1, X_2, \ldots, X_n$, их весов $w_1, w_2, \ldots, w_n$, а также внешних параметров δ и c

Require: $n > 0$ **and** $\forall i \in \{1, \ldots, n\} : |X_i| > 0$

$R_n \leftarrow R^{(n)}(X_1, X_2, \ldots, X_n, w_1, w_2, \ldots, w_n)$ {интегрированный результат на шаге n}
$W_n \leftarrow \sum_{i=1}^{n} w_i$ {суммарный вес наблюдений на шаге n}
$\hat{\Delta}_n \leftarrow \delta$
for $i = 1$ to n **do**
    $R_{n+1} \leftarrow R^{(2)}(R_n, X_i, W_n, w_i)$
    $\hat{\Delta}_n \leftarrow \hat{\Delta}_n + \rho_L(R_{n+1}, R_n)$
end for
$\hat{\Delta}_n \leftarrow \hat{\Delta}_n/(n+1)$
if $\hat{\Delta}_n \leqslant c$ **then**
    return ОСТАНОВ
else
    return ПРОДОЛЖИТЬ ПРОЦЕСС
end if

4.5 Экспериментальные результаты

В данном разделе будут продемонстрированы результаты экспериментального исследования метода останова процесса распознавания объекта в видеопотоке, представленного в разделе 4.4. В качестве объекта распознавания рассматривается текстовое поля документа, удостоверяющего личность. В простейшем случае результат распознавания текстового поля (элемент множества $\mathbb{X}$) можно представить виде строки над некоторым фиксированным алфавитом.

Для того, чтобы применить модель, описанную в разделе 4.2, необходимо определить метрику ρ и функции интеграции $R^{(n)}$ для множества строк $\mathbb{X}$. В качестве метрики на множестве строк, также как и в экспериментальном исследовании, представленном в главе 3 (см. раздел 3.5), предлагается использовать нормализованное расстояние Левенштейна [134]:

$$\rho_L(x,y) \stackrel{\text{def}}{=} \frac{2 \cdot \text{levenshtein}(x,y)}{|x| + |y| + \text{levenshtein}(x,y)}, \tag{4.20}$$

где $|x|$ — длина строки x, и $\text{levenshtein}(x,y)$ — расстояние Левенштейна между строками x и y. Значения этой метрики заключены в отрезке $[0, 1]$ и ее нормализация выполнена с сохранением неравенства треугольника.

В качестве функций интеграции $R^{(n)}$ использовался алгоритм ROVER [87], описание которого приведено в разделе 3.3. Для имплементации метода требуется ввести порог θ оценки пустого класса, учитывающийся в модуле голосования. В проведенных экспериментах, так же как и в экспериментальной части главы 3 (см. раздел 3.5), использовалось значение порога $\theta = 0.6$.

Экспериментальное исследование проводилось на открытом пакете данных MIDV-500 [88], содержащем видеоролики 50 документов, удостоверяющих личность, различных типов (по 10 видеороликов для каждого документа, по 30 кадров в видеоролике) с размеченными идеальными позициями и значениями текстовых полей. Были проанализированы 4 группы полей: даты, записанные цифрами и знаками препинания, номер документа, строки машиночитаемой зоны (MRZ, Machine-Readable Zone) и компоненты имени держателя документы, записанные латинским алфавитом.

Рассматривались только кадры, на которых документ целиком присутствует в кадре (следовательно видеопоследовательности в рассматриваемом

подмножестве пакета данных имели разную длину, от 1 до 30 кадров). Для того, чтобы минимизировать эффекты нормализации и обеспечить более ясное представление результатов, каждый клип был дополнен до 30 кадров путем повторения клипа с начала (таким образом, все анализируемые клипы имели одну и ту же длину 30).

Каждое поле вырезалось из исходного изображения при помощи проективного преобразования, согласно совместной разметке идеальных границ документа и координат текстового поля, с добавленными отступами, равными 10% от наименьшей стороны текстового поля. Размер вырезаемых изображений текстовых полей соответствовал разрешению 300 точек на дюйм. Каждое вырезанное текстовое поле распознавалось при помощи библиотеки распознавания с открытым исходным кодом Tesseract (версии 3.05.01 и 4.0.0) [138] используя параметры по умолчанию для английского языка. Все сравнения значений символов проводились вне зависимости от регистра, а также латинская буква «O» считалась идентичной цифре «0».

В таблице 4 для каждой группы текстовых полей приведены количество уникальных полей в пакете данных MIDV-500, общее количество изображений текстовых полей (среди всех кадров, на которых документ целиком присутствует в кадре), а также средняя длина последовательности кадров. Таблица 4 также приводит среднее расстояние между результатом X_i распознавания одиночного кадра и истинным значением X^*, и между интегрированным результатом для видеоролика и истинным значением X^* перед дополнением (R_{last}) и после дополнения (R_{30}), в терминах метрики ρ_L (4.20).

На рисунке 4.2 проиллюстрированы средние расстояния, по метрике ρ_L (4.20), от интегрированных результатов распознавания текстовой строки в видеопотоке до истинного значения, при анализе использовались все текстовые поля. Можно заметить, что ошибка значительно уменьшается во времени для всех групп полей, что может рассматриваться как практическое обоснование предположения (4.18).

На рисунке 4.3 продемонстрировано убывание разницы между расстояниями от соседних интегрированных результатов распознавания до истинного значения ($\mathrm{E}(\rho(R_n,X^*)) - \mathrm{E}(\rho(R_{n+1},X^*))$) во времени, среднего расстояния Δ_n между соседними интегрированными результатами распознавания и его оценки $\hat{\Delta}_n$ (4.19). В проведенных экспериментах использовалось значение настраиваемого параметра (4.19) $\delta = 0.2$. Хотя правило останова N_B (4.17) является

Таблица 4 — Средние значения метрики ρ_L до истинных значений для результатов распознавания при помощи библиотеки Tesseract [138] текстовых полей пакета данных MIDV-500 [88]. X_i – результат распознавания одиночного кадра, R_{last} – интегрированный результат распознавания видеоролика, полученный при помощи модификации алгоритма ROVER, R_{30} – интегрированный результат распознавания дополненного видеоролика, полученный при помощи модификации алгоритма ROVER

	Дата	Номер документа	Строки MRZ	Имя (латиница)	Все поля
Уникальных полей	91	48	30	79	248
Всего клипов	824	436	260	719	2239
Всего изображений	17735	9329	5096	15587	47747
Средняя длина клипа	21.523	21.397	19.600	21.679	21.325
Tesseract v3.05.01:					
$\mathrm{E}(\rho_L(X_i, X^*))$	0.360	0.422	0.258	0.443	0.388
$\mathrm{E}(\rho_L(R_{\text{last}}, X^*))$	0.244	0.326	0.162	0.338	0.281
$\mathrm{E}(\rho_L(R_{30}, X^*))$	0.246	0.323	0.164	0.336	0.280
Tesseract v4.0.0:					
$\mathrm{E}(\rho_L(X_i, X^*))$	0.238	0.287	0.339	0.250	0.262
$\mathrm{E}(\rho_L(R_{\text{last}}, X^*))$	0.123	0.160	0.277	0.125	0.149
$\mathrm{E}(\rho_L(R_{30}, X^*))$	0.125	0.163	0.279	0.127	0.151

грубой аппроксимацией близорукого правила N_A (4.12), однако можно отметить как практическую обоснованность предположения (4.13), так и обоснованность оценки $\hat{\Delta}_n$ (4.19) начиная с шага $n = 2$, для целей поставленной задачи. Аппроксимацию правила останова N_B теперь будем строить при помощи порогового отсечения оценки $\hat{\Delta}_n$.

Для того, чтобы оценить эффективность правила останова может быть построен профиль эффективности, графически показывающий зависимость среднего количества обработанных наблюдений и соответствующего среднего расстояния от полученного интегрированного результата в момент останова до истинного значения, варьируя стоимость наблюдения c. Подобный профиль эффективности отражает размен между временем, необходимым для обработки видеопоследовательности, и точностью полученного результата распознавания, а также позволяет визуально сравнить различные стратегии останова.

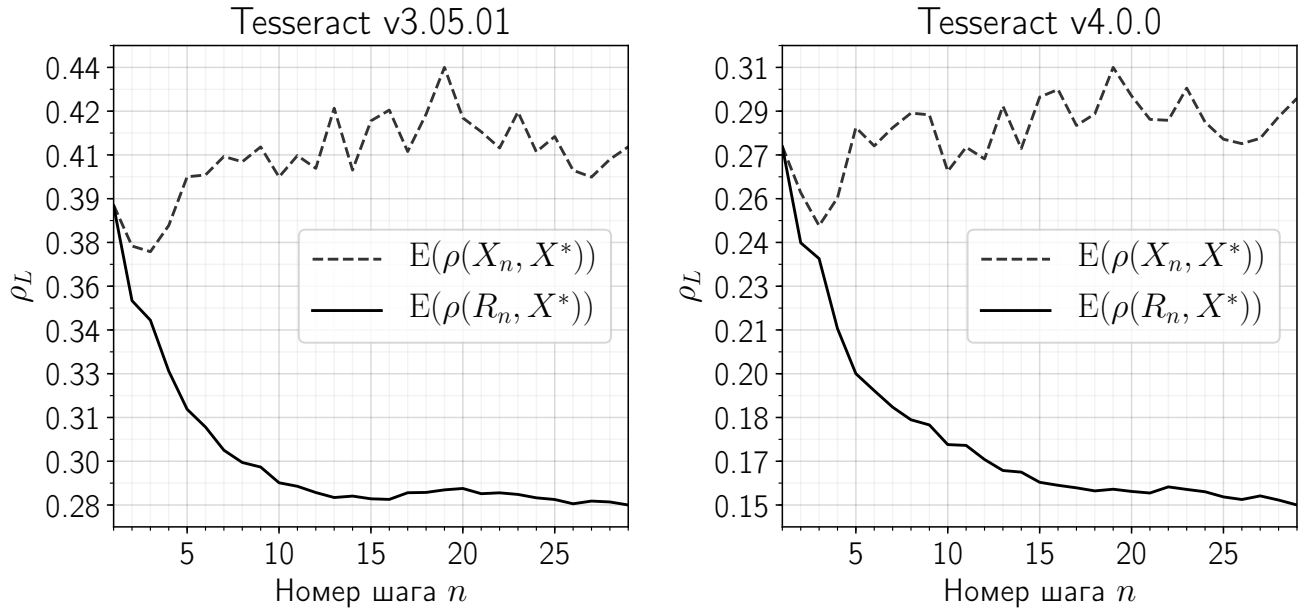


Рисунок 4.2 — Средние расстояния от покадрового результата распознавания текстовой строки и от интегрированного результата распознавания в видеопотоке до истинного значения, по метрике ρ_L (4.20). Распознавание текстовых полей производилось при помощи библиотеки Tesseract v3.05.01 (слева) и v4.0.0 (справа)

В качестве контрольного правила использовалось простое правило подсчета N_K, которое требует останавливать процесс распознавания на фиксированном шаге K. Дополнительно исследовались два варианта правила останова, описанных в [139]. Поскольку оригинальная работа опирается на использование показателей уверенности результата распознавания, которые недоступны в рамках исследуемой в этой главе модели, правило останова, описанное в [139], вырождается в пороговое отсечение размера наибольшего кластера идентичных результатов распознавания, накопленных к моменту n. Таким образом, построено два контрольных правила останова: N_{CX}, производящий пороговое отсечение размера наибольшего кластера идентичных результатов покадрового распознавания $x_1, \ldots, x_n$, и N_{CR}, аналогично рассматривающий интегрированные результаты распознавания $R_1, \ldots, R_n$. Наконец, правило останова N_B (4.17), построенное в данной главе, оценивающее на шаге n ожидаемое расстояние Δ_n до следующего интегрированного результата и останавливающее процесс в момент, когда эта оценка становится меньше или равной порогу. Правило останова N_B действует только начиная с шага $n = 2$ (т.е. с шага, на котором оценка $\hat{\Delta}_n$ (4.19) становится более обоснованной).

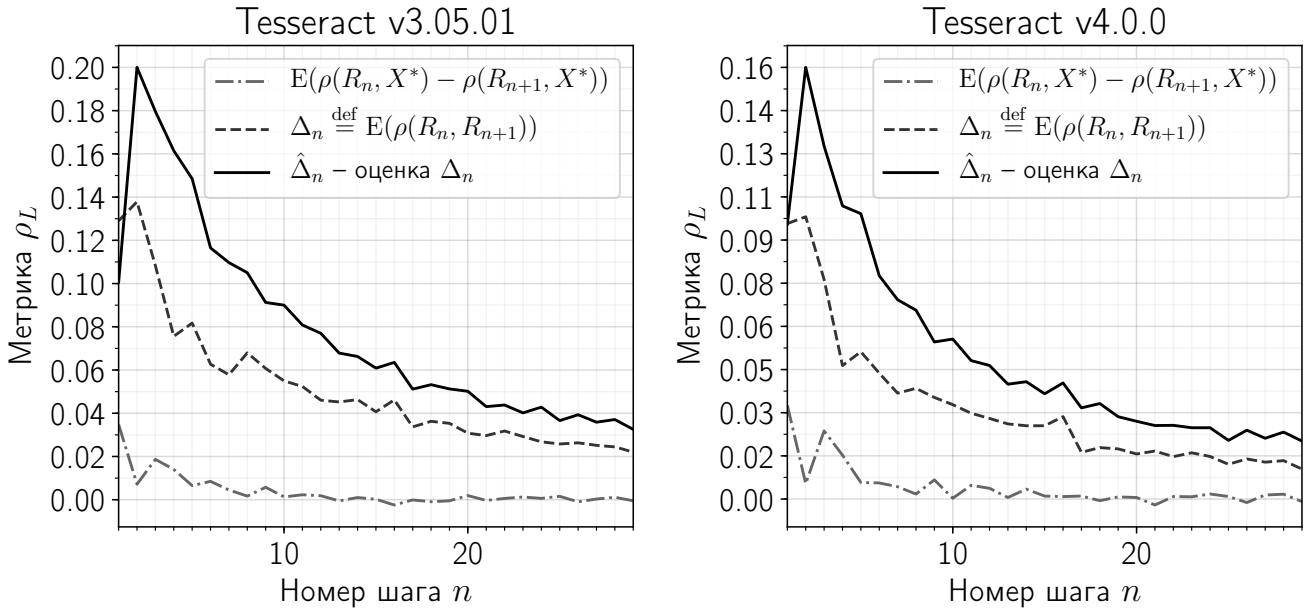


Рисунок 4.3 — Убывание среднего расстояния между соседними интегрированными результатами распознавания и его оценка, при значении настраиваемого параметра $\delta = 0.2$. Распознавание текстовых полей производилось при помощи библиотеки Tesseract v3.05.01 (слева) и v4.0.0 (справа)

Рисунок 4.4 иллюстрирует эффективность правил останова для всех групп текстовых полей, распознаваемых при помощи библиотеки Tesseract (версии 3.05.01 и 4.0.0). Более низкое положение кривой отражает большую эффективность правила останова. Можно отметить, что с среднем предлагаемое правило останова N_B (4.17) обладает большей эффективностью, чем другие исследованные методы. Стоит отметить, что метод останова N_B (4.17) обладает высокой эффективностью без каких-либо модификаций для двух различных версий библиотеки Tesseract, использующих различные поколения алгоритмов распознавания текстовой строки.

В таблице 5 показано среднее расстояние от интегрированного результата до правильного ответа в момент останова, которое может быть достигнуто при помощи исследованных правил останова, при распознавании текстовых полей при помощи библиотеки Tesseract v3.05.01. Колонки таблицы 5 отражают целевые интервалы для значений среднего количество использованных наблюдений (т.е. среднего количества обработанных кадров), строки таблицы соответствуют правилам останова, и каждая ячейка содержит результат замера с наименьшим средним количеством наблюдений, попадающим в данный интервал. Некоторые ячейки таблицы не содержат данных (помечены символом $\varnothing$) – это означает, что

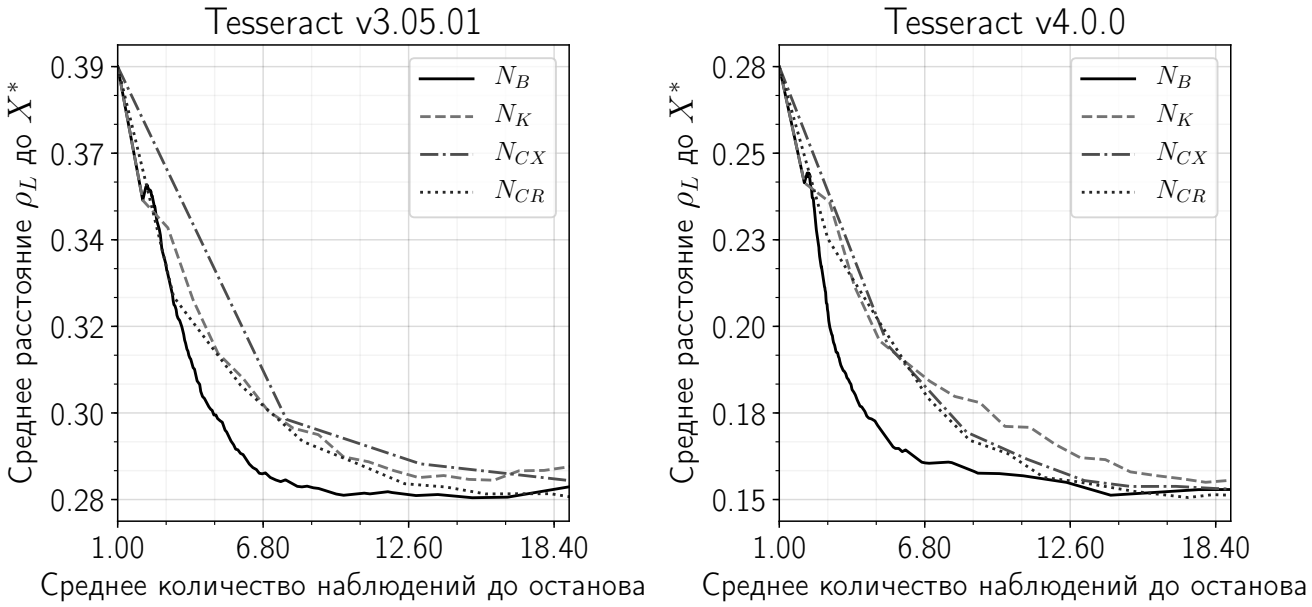


Рисунок 4.4 — Сравнительное исследование эффективности правил останова: график зависимости среднего расстояния между полученным результатом в момент останова и истинным значением от среднего количества обработанных кадров до останова, при изменяющейся стоимости наблюдения c, при значении настраиваемого параметра $\delta = 0.2$. Распознавание текстовых полей производилось при помощи библиотеки Tesseract v3.05.01 (слева) и v4.0.0 (справа)

соответствующее правило останова не способно достичь среднего количества обработанных кадров в соответствующем интервале для рассматриваемого набора входных данных (поскольку обладает более дискретной природой). Можно сделать вывод, что почти во всех целевых интервалах правило останова N_B (4.17) показывает лучший результат среди исследуемых альтернатив. Подобный результат также наблюдается при распознавании текстовых полей при помощи библиотеки Tesseract v4.0.0 (результаты представлены в таблице 6).

Профили эффективности правил останова для отдельных групп полей (Дата, Номер документа, MRZ строка, Имя (латиница)) представлены на рисунке 4.5.

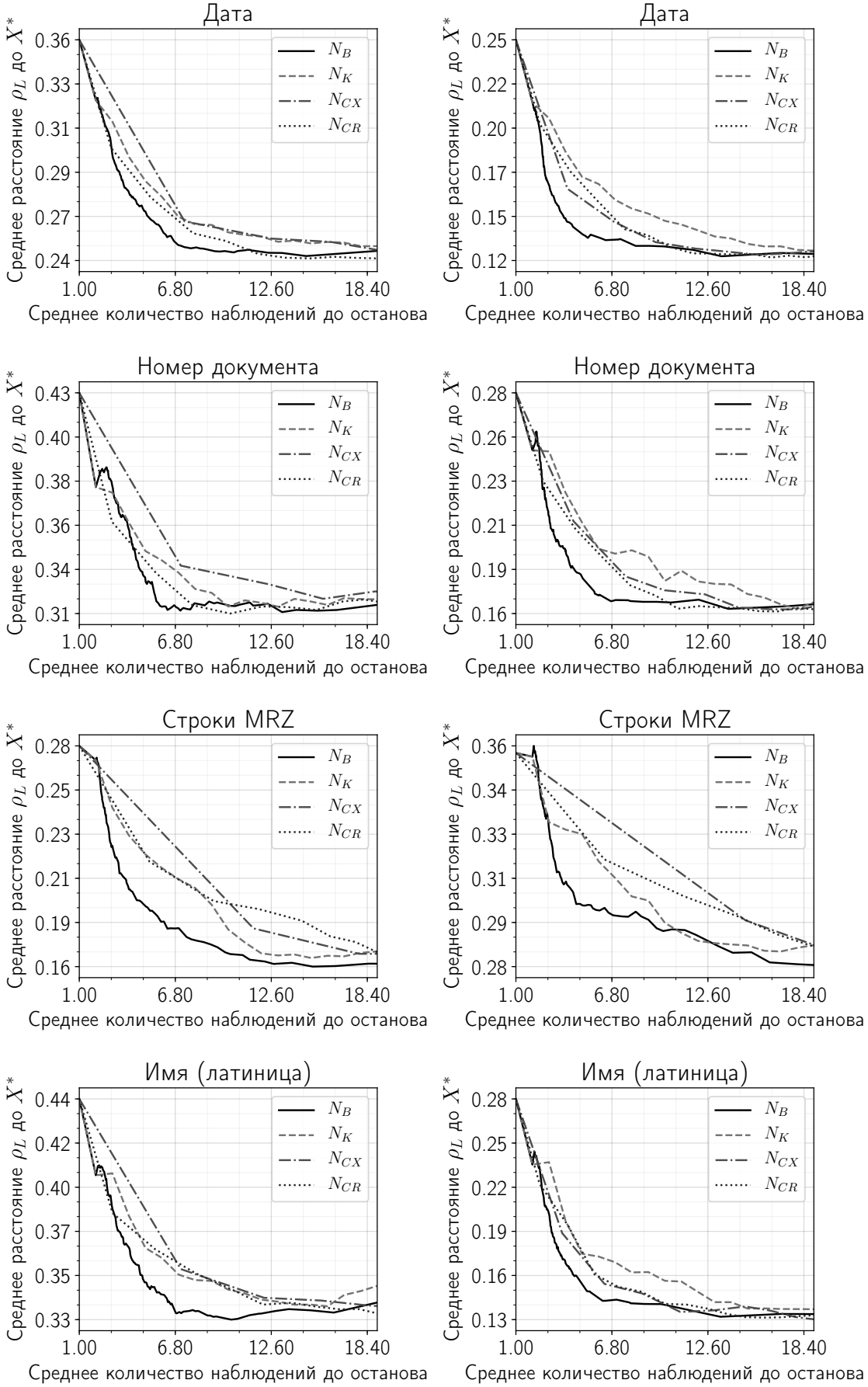


Рисунок 4.5 — Профили эффективности правил останова, для различных групп полей. Распознавание текстовых полей производилось при помощи библиотеки Tesseract v3.05.01 (слева) и v4.0.0 (справа)

Таблица 5 — Достигнутые значения среднего расстояния от интегрированного результата до идеального значения в момент останова, в терминах метрики ρ_L, распознавание проводилось при помощи Tesseract v3.05.01

Правило останова	Измеряемый параметр	Целевой интервал среднего количества наблюдений $E(N)$								
		3 ± 0.5	4 ± 0.5	5 ± 0.5	6 ± 0.5	7 ± 0.5	8 ± 0.5	9 ± 0.5	10 ± 0.5	11 ± 0.5
N_{CX}	$\mathrm{E}(N)$ $\mathrm{E}(\rho_L(R_N, X^*))$	∅	∅	∅	∅	∅	7.727 0.298	∅	∅	∅
N_{CR}	$\mathrm{E}(N)$ $\mathrm{E}(\rho_L(R_N, X^*))$	3.230 **0.329**	∅	∅	5.909 0.306	∅	8.375 0.292	∅	∅	**10.560** 0.286
N_K	$\mathrm{E}(N)$ $\mathrm{E}(\rho_L(R_N, X^*))$	3.000 0.347	4.000 0.329	5.000 0.315	6.000 0.308	7.000 0.300	8.000 0.295	9.000 0.294	10.000 0.288	11.000 0.287
$\boldsymbol{N_B}$	$\mathrm{E}(N)$ $\mathrm{E}(\rho_L(R_N, X^*))$	**2.509** 0.351	**3.558** **0.322**	**4.527** **0.302**	**5.521** **0.292**	**6.531** **0.285**	**7.691** **0.282**	**8.554** **0.280**	**9.715** **0.278**	10.830 **0.278**

Таблица 6 — Достигнутые значения среднего расстояния от интегрированного результата до идеального значения в момент останова, в терминах метрики ρ_L, распознавание проводилось при помощи Tesseract v4.0.0

Правило останова	Измеряемый параметр	Целевой интервал среднего количества наблюдений $E(N)$								
		3 ± 0.5	4 ± 0.5	5 ± 0.5	6 ± 0.5	7 ± 0.5	8 ± 0.5	9 ± 0.5	10 ± 0.5	11 ± 0.5
N_{CX}	$\mathrm{E}(N)$ $\mathrm{E}(\rho_L(R_N, X^*))$	∅	∅	5.332 0.195	∅	∅	8.471 0.170	∅	∅	10.901 0.162
N_{CR}	$\mathrm{E}(N)$ $\mathrm{E}(\rho_L(R_N, X^*))$	2.936 0.227	∅	5.099 0.201	∅	6.920 0.180	∅	**8.594** 0.167	10.103 0.164	∅
N_K	$\mathrm{E}(N)$ $\mathrm{E}(\rho_L(R_N, X^*))$	3.000 0.237	4.000 0.213	5.000 0.197	6.000 0.191	7.000 0.185	8.000 0.180	9.000 0.178	10.000 0.171	11.000 0.171
$\boldsymbol{N_B}$	$\mathrm{E}(N)$ $\mathrm{E}(\rho_L(R_N, X^*))$	**2.580** **0.224**	**3.551** **0.188**	**4.571** **0.174**	**5.539** **0.165**	**6.683** **0.161**	**7.742** **0.161**	8.771 **0.158**	**9.779** **0.158**	**10.726** **0.157**

4.6 Выводы по главе

В главе была рассмотрена задача останова процесса распознавания объекта в видеопотоке, что является важной и новой задачей, особенно актуальной при разработке систем оптического распознавания, предназначенных для работы на мобильном устройстве. Была предложена формальная постановка задачи, следующая классической формулировке задачи останова и предложен метод, рассматривающий процесс распознавания объекта в видеопотоке как процесс, останов в котором становится монотонным начиная с определенного шага.

На основе предложенного метода был разработан алгоритм останова процесса распознавания строкового объекта в видеопотоке, в котором оценка расстояния между текущим и следующим интегрированными результатами вычисляется путем моделирования следующего интегрированного результата с использованием уже накопленных наблюдений.

Метод был экспериментально апробирован в задаче распознавания текстовых полей документов, удостоверяющих личность, на открытом пакете данных MIDV-500 и с использованием широко доступной библиотеки распознавания текстовых полей Tesseract с открытым исходным кодом. Было продемонстрировано, что предложенное правило останова является более эффективным, чем пороговое отсечение количества обработанных кадров или пороговое отсечение размеров максимального кластера идентичных результатов, несмотря на то, что в модели не участвуют оценки уверенности результатов распознавания.

Заключение

Основные результаты работы заключаются в следующем.

1. Построена математическая модель системы распознавания объекта в видеопотоке с блоком комбинирования покадровых результатов распознавания и с блоком принятия решения об останове. В качестве функционала эффективности системы была рассмотрена линейная комбинация расстояния от интегрированного результата распознавания до истинного значения объекта и штрафной функции от времени от момента начала процесса съемки до останова. Данная модель позволяет рассматривать систему распознавания объекта в видеопотоке как итерационный вычислительный процесс, который способен выдать в любое время наилучшее на данный момент решение, и прекратить захват новых изображений согласно заданному правилу останова.

2. Выполнено оригинальное исследование влияния характеристик входных данных на выбор оптимальной стратегии комбинирования покадровых результатов классификации в рамках задачи распознавания одиночного символа в видеопотоке. Показано, что если в последовательности обрабатываемых изображениях одиночного изображения отсутствуют ошибки предварительной обработки (такие, как ошибки локализации и сегментации символов), более высокую точность финального результата обеспечивает правило максимальной оценки. Для видеопоследовательностей, в которых встречаются ошибки локализации и сегментации символов, более высокую точность финального результата обеспечивают правила произведения оценок, правило голосования и правило суммы оценок.

3. Разработан новый алгоритм комбинирования результатов распознавания строкового объекта, учитывающий альтернативные варианты классификации отдельных символов (компонентов строкового объекта). Экспериментально показано, что предложенный алгоритм способен обеспечить более высокую точность интегрированного результата по сравнению с методом интеграции результатов распознавания как строк над множеством классов значений компонентов, применительно к задаче распознавания текстовой строки в видеопотоке.

4. Была рассмотрена задача останова процесса распознавания объекта в видеопотоке, что является важной и новой задачей, особенно актуальной при разработке систем оптического распознавания, предназначенных для работы на

Таблица 7 — Достигнутые наилучшие значения среднего расстояния от интегрированного результата до идеального значения в момент останова; результаты распознавания интегрированы при помощи Алгоритма 1

Метод останова	Наилучшая точность при ограничении среднего числа кадров					
	$\leqslant 3$	$\leqslant 4$	$\leqslant 5$	$\leqslant 6$	$\leqslant 7$	$\leqslant 8$
N_{CX}	$\varnothing$	0.083	0.080	0.078	0.073	0.072
N_{CR}	0.096	0.084	0.080	0.077	0.074	0.072
N_K	0.115	0.104	0.097	0.089	0.084	0.082
Алг. 2	**0.092**	**0.082**	**0.076**	**0.073**	**0.072**	**0.070**

мобильных устройствах. Разработан новый метод останова процесса распознавания объекта в видеопотоке на основе порогового отсечения оценки ожидаемого расстояния между текущим и следующим интегрированными результатами. Метод разработан исходя из предположения о том, что задача останова процесса распознавания становится монотонной начиная с некоторого шага. На основе разработанного метода предложен новый алгоритм останова процесса распознавания строкового объекта в видеопотоке, в котором оценка вычисляется путем моделирования следующего интегрированного результата с использованием уже накопленных наблюдений. Было продемонстрировано, что в задаче распознавания текстовых строк предложенное правило останова является более эффективным, чем пороговое отсечение количества обработанных кадров или пороговое отсечение размеров максимального кластера идентичных результатов.

5. Совместное использование разработанных алгоритмов (Алгоритм 1 комбинирования результатов распознавания строковых объектов, учитывающий альтернативные варианты классификации символов, и Алгоритм 2 останова процесса распознавания строки) позволяет достичь большей точности распознавания при том же среднем количестве обработанных изображений. В таблице 7 показаны достигнутые наилучшие значения среднего расстояния от результата до истинного значения при различных ограничениях на среднее количество обработанных кадров, с интегрирование результатов Алгоритмом 1 и при использовании рассмотренных алгоритмов останова, включая Алгоритм 2.

6. Результаты работы в качестве программных компонентов систем распознавания документов в видеопотоке были внедрены в программное обеспечение «Smart 3D OCR MRZ» и «Smart PassportReader» компании ООО «Смарт

Энджинс РУС», а также «Smart IDReader» компании ООО «Смарт Энджинс Сервис». Данные продукты интегрированы в информационную инфраструктуру ряда коммерческих организаций, а также в ряд информационных решений государственных структур Российской Федерации.

Основные результаты по теме диссертации изложены в 14 публикациях, в том числе: 6 изданы в журналах, рекомендованных ВАК, 3 – в сборниках трудов конференций (входящих в международные базы цитирования Scopus и Web of Science), 2 патента на полезную модель и 3 свидетельства о государственной регистрации программы для ЭВМ.

Список литературы

1. Национальный стандарт РФ ГОСТ Р 7.0.8-2013 «Система стандартов по информации, библиотечному и издательскому делу. Делопроизводство и архивное дело. Термины и определения». — М. : Стандартинформ, 2014. — 16 с.
2. Федеральный закон 77-ФЗ «Об обязательном экземпляре документов» от 29.12.1994. — URL: http://www.consultant.ru/document/cons_doc_LAW_5437 (дата обр. 25.06.2017).
3. *Schantz H. F.* History of OCR, Optical Character Recognition. — Recognition Technologies Users Association, 1982. — 114 p.
4. *Palmer R.* The Bar Code Book: A Comprehensive Guide to Reading, Printing, Specifying, Evaluating, and Using Bar Code and Other Machine-readable Symbols. — Trafford Publishing, 2007. — 470 p.
5. *Шоломов Д. Л.* Синтаксические методы контекстной обработки в задачах распознавания текста. — Автореф. дис. ... канд. тех. наук. — М. : Институт системного анализа РАН, 2007.
6. *Арлазаров В. В.* Структурирование визуальных представлений информационной среды и методы определения надежности распознавания. — Автореф. дис. ... канд. тех. наук. — М. : Московский государственный институт стали и сплавов (технологический университет), 2004.
7. *Арлазаров В. В.*, *Булатов К. Б.*, *Карпенко С. М.* Метод определения надежности распознавания в задаче распознавания тисненых символов // Труды ИСА РАН. — 2013. — Т. 63, № 3. — С. 117—122.
8. Radium One: Mobile Marketing Survey Report [Электронный ресурс]. — 2014. — URL: http : / / cfile219 . uf . daum . net / attach / 237B5C39545D58DD2F4892 (дата обр. 25.06.2017).
9. Litera Corp: Mobile Device Usage and Document Security Survey Results [Электронный ресурс]. — 2013. — URL: http://www.litera.com/wp-content/uploads/2015/12/Mobile-Device-Usage-Survey-Results-2013_final.pdf (дата обр. 25.06.2017).

10. 32% of UK consumers make purchases on a smartphone: stats [Электронный ресурс]. — 2014. — URL: https://econsultancy.com/blog/64511-32-of-uk-consumers-make-purchases-on-a-smartphone-stats#i.w36j8hi93emrpd (дата обр. 25.06.2017).

11. Федеральный закон 152-ФЗ «О персональных данных» от 27.07.2006. — URL: http://www.consultant.ru/document/cons_doc_LAW_61801 (дата обр. 25.06.2017).

12. Федеральный закон 115-ФЗ «О противодействии легализации (отмыванию) доходов, полученных преступным путем, и финансированию терроризма» от 07.08.2001 (ред. от 29.07.2017). — URL: http://www.consultant.ru/document/cons_doc_LAW_32834 (дата обр. 07.01.2018).

13. Положение Банка России N 499-П «Об идентификации кредитными организациями клиентов, представителей клиента, выгодоприобретателей и бенефициарных владельцев в целях противодействия легализации (отмыванию) доходов, полученных преступным путем, и финансированию терроризма» (с изменениями и дополнениями) от 15.10.2015. — URL: http://base.garant.ru/71277312 (дата обр. 07.01.2018).

14. *Hsueh M.* Interactive Text Recognition and Translation on a Mobile Device : tech. rep. / EECS Department, University of California, Berkeley. — 2011. — UCB/EECS-2011–57. — URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2011/EECS-2011-57.html (visited on 06/25/2017).

15. Анализ особенностей использования стационарных и мобильных малоразмерных цифровых видео камер для распознавания документов / В. В. Арлазаров [и др.] // Информационные технологии и вычислительные системы. — 2014. — № 3. — С. 71—78.

16. Skew Estimation by Instances / S. Uchida [et al.] // Proceedings of the 2008 The Eighth IAPR International Workshop on Document Analysis Systems. — Washington, DC, USA : IEEE Computer Society, 2008. — P. 201–208. — (DAS '08).

17. *Ishitani Y.* Document skew detection based on local region complexity // Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on. — 1993. — P. 49–52.

18. *Lu Y.*, *Tan C. L.* Improved nearest neighbor based approach to accurate document skew estimation // Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. Vol. 1. — 2003. — P. 503–507.

19. *Lu Y.*, *Tan C. L.* Camera document restoration for OCR // Proceedings of the 1st International Workshop on Camera-Based Document Analysis and Recognition, CBDAR 2005. — 2005. — P. 17–24.

20. *Pratt W. K.* Digital Image Processing: PIKS Scientific Inside. — 4rd. — Locas Altos, California, USA : PixelSoft, Inc., 2007. — 807 p.

21. *Hinds S. C.*, *Fisher J. L.*, *D'Amato D. P.* A document skew detection method using run-length encoding and the Hough transform // Proceedings. 10th International Conference on Pattern Recognition. Vol. 1. — 1990. — P. 464–468.

22. *Le D. S.*, *Thoma G. R.*, *Wechsler H.* Automated Page Orientation and Skew Angle Detection for Binary Document Images // Pattern Recognition. — 10/1994. — P. 1325–1344.

23. *Yu B.*, *Jain A. K.* A Robust and Fast Skew Detection Algorithm for Generic Documents // Pattern Recognition. Vol. 29. — 1996. — P. 1599–1629.

24. Hough transform: underestimated tool in the computer vision field / D. P. Nikolaev [et al.] // Proceedings of the 22th European Conference on Modelling and Simulation. Vol. 238. — 2008. — P. 238–246.

25. *Safabakhsh R.*, *Khadivi S.* Document skew detection using minimum-area bounding rectangle // Proceedings International Conference on Information Technology: Coding and Computing (Cat. No.PR00540). — 2000. — P. 253–258.

26. *Jain R.*, *Kasturi R.*, *Schunck B.* Machine Vision. — McGraw-Hill, 1995. — 549 p. — (Computer Science Series).

27. *Clark P.*, *Mirmehdi M.* Rectifying perspective views of text in 3D scenes using vanishing points // Pattern Recognition. — 2003. — Vol. 36, no. 11. — P. 2673–2686.

28. *Dance C. R.* Perspective estimation for document images // Proc. SPIE. Vol. 4670. — 2001. — P. 244–254.

29. *Lu S.*, *Chen B. M.*, *Ko C.* Perspective rectification of document images using fuzzy set and morphological operations // Image and Vision Computing. — 2005. — Vol. 23, no. 5. — P. 541–553.

30. *Pilu M.* Extraction of illusory linear clues in perspectively skewed documents // Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. Vol. 1. — 2001. — P. I-363–I-368.

31. Segments Graph-Based Approach for Document Capture in a Smartphone Video Stream / A. Zhukovsky [et al.]. — 2017.

32. *Shufelt J. A.* Performance evaluation and analysis of vanishing point detection techniques // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 1999. — Vol. 21, no. 3. — P. 282–288.

33. *Castleman K.* Digital Image Processing. — Prentice Hall, 1996. — 667 p. — (Prentice-Hall signal processing series).

34. *Hartley R.*, *Zisserman A.* Multiple View Geometry in Computer Vision. — Cambridge University Press, 2003. — 655 p. — (Cambridge books online).

35. Perspective rectification for mobile phone camera-based documents using a hybrid approach to vanishing point detection / X.-C. Yin [et al.] // Proceedings of the 2nd International Workshop on Camera-Based Document Analysis and Recognition, CBDAR 2007. — 2007. — P. 37–44.

36. Rectification and Recognition of Text in 3-D Scenes / G. K. Myers [et al.] // Int. J. Doc. Anal. Recognit. — Berlin, Heidelberg, 2005. — Vol. 7, no. 2/3. — P. 147–158.

37. *Viola P.*, *Jones M. J.* Robust Real-Time Face Detection // Int. J. Comput. Vision. — Hingham, MA, USA, 2004. — Vol. 57, no. 2. — P. 137–154.

38. Visual appearance based document image classification / S. Usilin [et al.] // 2010 IEEE International Conference on Image Processing. — 2010. — P. 2133–2136.

39. Generalization of the Viola-Jones method as a decision tree of strong classifiers for real-time object recognition in video stream / A. Minkina [et al.] // Proc. SPIE. Vol. 9445. — 2015. — P. 944517-944517–5.

40. *Арлазаров В.*, *Булатов К.*, *Чернов Т.* Метод нечеткого поиска изображений в больших объемах видеоданных // Системы высокой доступности. — 2016. — Т. 12, № 1. — С. 53—58.

41. A document straight line based segmentation for complex layout extraction / F. Cloppet [et al.]. — 2017.

42. *Melinda L.* Document Layout Analysis using Multigaussian Fitting. — 2017.

43. Page Segmentation for Historical Handwritten Documents Using Fully Convolutional Networks / Y. Xu [et al.]. — 2017.

44. A Robust and Binarization-Free Approach for Text Line Detection in Historical Documents / T. Gr [et al.]. — 2017.

45. *Moysset B.*, *Kermorvant C.*, *Wolf C.* Full-Page Text Recognition: Learning Where to Start and When to Stop. — 2017. — eprint: 1704.08628.

46. Text Localization in Natural Images Using Stroke Feature Transform and Text Covariance Descriptors / W. Huang [et al.] // 2013 IEEE International Conference on Computer Vision. — 2013. — P. 1241–1248.

47. *Gaddour H.*, *Kanoun S.*, *Vincent N.* Color Stability and Homogeneity Regions to Detect Text in Real Scene Images : CSHR. — 2017.

48. *Turki H.*, *Halima M. B.*, *Alimi A. M.* Text Detection based on MSER and CNN Features. — 2017.

49. A Robust Symmetry-based Method for Scene / Video Text Detection Through Neural Network / Y. Wu [et al.]. — 2017.

50. Max-Pooling based Scene Text Proposal for Scene Text Detection / D. N. Van [et al.]. — 2017.

51. *Qin S.*, *Manduchi R.* Cascaded Segmentation-Detection Networks for Word-Level Text Spotting. — 2017. — eprint: 1704.00834.

52. Text Detection by Faster R-CNN with Multiple Region Proposal Networks / Y. Nagaoka [et al.]. — 2017.

53. Deep Residual Text Detection Network for Scene Text / X. Zhu [et al.]. — 2017. — eprint: 1711.04147.

54. *Casey R. G.*, *Lecolinet E.* A Survey of Methods and Strategies in Character Segmentation // IEEE Trans. Pattern Anal. Mach. Intell. — Washington, DC, USA, 1996. — Vol. 18, no. 7. — P. 690–706.

55. *Saba T.*, *Sulong G.*, *Rehman A.* A survey on methods and strategies on touched characters segmentation // International Journal of Research and Reviews in Computer Science. — 2010. — Vol. 1, no. 2. — P. 103–114.

56. Исследование методов сегментации изображений текстовых блоков документов с помощью алгоритмов структурного анализа и машинного обучения / Т. С. Чернов [и др.] // Вестник РФФИ. Обработка изображений и распознавание образов. — 2016. — Т. 92, № 4. — С. 55—71.

57. Grayscale-projection based Optimal Character Segmentation for Camera-captured Faint Text Recognition / F. Jia [et al.]. — 2017.

58. Combining Convolutional Neural Networks and LSTMs for Segmentation-Free OCR / S. Rawls [et al.]. — 2017.

59. *Breuel T. M.* High Performance Text Recognition using a Hybrid Convolutional-LSTM Implementation. — 2017.

60. *Горелик А. Л.*, *Скрипкин В. А.* Методы распознавания: Учебное пособие. 2-е издание, переработанное и дополненное. — М. : Высшая школа, 1984. — 208 с.

61. DeepFace: Closing the Gap to Human-Level Performance in Face Verification / Y. Taigman [et al.] // 2014 IEEE Conference on Computer Vision and Pattern Recognition. — 2014. — P. 1701–1708.

62. *Славин О. А.* Адаптивное распознавание и его применение к системе ввода печатного текста : дис. ... докт. / Славин О. А. — М. : Институт системного анализа РАН, 2011.

63. Gradient-Based Learning Applied to Document Recognition / Y. LeCun [et al.] // Proceedings of the IEEE. — 1998.

64. *Krizhevsky A.*, *Sutskever I.*, *Hinton G. E.* ImageNet Classification with Deep Convolutional Neural Networks // Advances in Neural Information Processing Systems 25 / ed. by F. Pereira [et al.]. — Curran Associates, Inc., 2012. — P. 1097–1105.

65. *Zeiler M. D.*, *Fergus R.* Visualizing and Understanding Convolutional Networks // Computer Vision – ECCV 2014 / ed. by D. Fleet [et al.]. — Cham : Springer International Publishing, 2014. — P. 818–833.

66. *Simonyan K.*, *Zisserman A.* Very Deep Convolutional Networks for Large-Scale Image Recognition // CoRR. — 2014. — Vol. abs/1409.1556.

67. Going deeper with convolutions / C. Szegedy [et al.] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2015. — P. 1–9.

68. Deep Residual Learning for Image Recognition / K. He [et al.] // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2016. — P. 770–778.

69. *Ronneberger O.*, *Fischer P.*, *Brox T.* U-Net: Convolutional Networks for Biomedical Image Segmentation // Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 / ed. by N. Navab [et al.]. — Cham : Springer International Publishing, 2015. — P. 234–241.

70. *Moosavi-Dezfooli S.*, *Fawzi A.*, *Frossard P.* DeepFool: a simple and accurate method to fool deep neural networks // CoRR. — 2015. — Vol. abs/1511.04599. — arXiv: 1511.04599. — URL: http://arxiv.org/abs/1511.04599 (visited on 06/03/2018).

71. The Limitations of Deep Learning in Adversarial Settings / N. Papernot [et al.] // CoRR. — 2015. — Vol. abs/1511.07528. — arXiv: 1511.07528. — URL: http://arxiv.org/abs/1511.07528 (visited on 06/03/2018).

72. *Su J.*, *Vargas D. V.*, *Sakurai K.* One pixel attack for fooling deep neural networks // CoRR. — 2017. — Vol. abs/1710.08864. — arXiv: 1710.08864. — URL: http://arxiv.org/abs/1710.08864 (visited on 06/03/2018).

73. Regularizing Neural Networks by Penalizing Confident Output Distributions / G. Pereyra [et al.] // CoRR. — 2017. — Vol. abs/1701.06548.

74. Post-Processing OCR Text Using Web-Scale Corpora / J. Mei [et al.] // Proceedings of the 2017 ACM Symposium on Document Engineering. — Valletta, Malta : ACM, 2017. — P. 117–120. — (DocEng '17).

75. *Hammarström H.*, *Virk S. M.*, *Forsberg M.* Poor Man's OCR Post-Correction: Unsupervised Recognition of Variant Spelling Applied to a Multilingual Document Collection // Proceedings of the 2Nd International Conference on Digital Access to Textual Cultural Heritage. — Göttingen, Germany : ACM, 2017. — P. 71–75. — (DATeCH2017).

76. *Bouchaffra D.*, *Govindaraju V.*, *Srihari S. N.* Postprocessing of Recognized Strings Using Nonstationary Markovian Models // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 1997. — Vol. 21, no. 10. — P. 990–999.

77. *Forney G. D.* The Viterbi Algorithm: A Personal History // CoRR. — 2005. — URL: http://arxiv.org/abs/cs/0504020 (дата обр. 26.06.2017).

78. *Kukich K.* Techniques for Automatically Correcting Words in Text // ACM computing survey: Computational Linguistic. — 1992. — Vol. 24, no. 4. — P. 377–439.

79. OCR Post-processing Using Weighted Finite-State Transducers / R. Llobet [et al.] // Proceedings of the 2010 20th International Conference on Pattern Recognition. — 2010. — P. 2021–2024.

80. *Hart P. E.*, *Nilsson N. J.*, *Raphael B.* A Formal Basis for the Heuristic Determination of Minimum Cost Paths // IEEE Transactions on Systems Science and Cybernetics. — 1968. — Vol. 4, no. 2. — P. 100–107.

81. A Man-Machine Cooperating System Based on the Generalized Reject Model / S. Kimura [et al.] // 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Vol. 01. — 2017. — P. 1324–1329.

82. *Doermann D.*, *Liang J.*, *Huiping L.* Progress in camera-based document image analysis // Proceedings of Seventh International Conference on Document Analysis and Recognition. — 2003. — Vol. 1. — P. 606–616.

83. Проблемы распознавания машиночитаемых зон с использованием малоформатных цифровых камер мобильных устройств / К. Б. Булатов [и др.] // Труды ИСА РАН. — 2015. — Т. 65, № 3. — С. 85—93.

84. *Chen D.* Text detection and recognition in images and video sequences : Master's thesis / Chen D. — Lausanne : EPFL, 2003. — 141 p. — Thesis 2863.

85. *Wemhoener D.*, *Yalniz I. Z.*, *Manmatha R.* Creating an Improved Version Using Noisy OCR from Multiple Editions // Proceedings of the 2013 12th International Conference on Document Analysis and Recognition. — Washington, DC, USA : IEEE Computer Society, 2013. — P. 160–164. — (ICDAR '13).

86. *Lopresti D.*, *Zhou J.* Using Consensus Sequence Voting to Correct OCR Errors // Comput. Vis. Image Underst. — New York, NY, USA, 1997. — Vol. 67, no. 1. — P. 39–47.

87. *Fiscus J. G.* A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER) // 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings. — 1997. — P. 347–354.

88. A Dataset for Identity Documents Analysis and Recognition on Mobile Devices in Video Stream / V. V. Arlazarov [et al.] // ArXiv e-prints. — 2018. — "arXiv": 1807.05786 (cs.CV).

89. *Rokach L.* Ensemble-based classifiers // Artificial Intelligence Review. — 2010. — Vol. 33, no. 1. — P. 1–39.

90. On Combining Classifiers / J. Kittler [et al.] // IEEE Trans. Pattern Anal. Mach. Intell. — Washington, DC, USA, 1998. — Vol. 20, no. 3. — P. 226–239.

91. *Rogova G.* Combining the Results of Several Neural Network Classifiers // Neural Netw. — Oxford, UK, UK, 1994. — Vol. 7, no. 5. — P. 777–781.

92. *Quost B.*, *Masson M.-H.*, *Denœux T.* Classifier Fusion in the Dempster–Shafer Framework Using Optimized T-norm Based Combination Rules // Int. J. Approx. Reasoning. — New York, NY, USA, 2011. — Vol. 52, no. 3. — P. 353–374.

93. *Ting K. M.*, *Witten I. H.* Issues in Stacked Generalization // J. Artif. Int. Res. — USA, 1999. — Vol. 10, no. 1. — P. 271–289.

94. *Kuncheva L. I.*, *Bezdek J. C.*, *Duin R. P.* Decision templates for multiple classifier fusion: an experimental comparison // Pattern Recognition. — 2001. — Vol. 34, no. 2. — P. 299–314.

95. *Merz C. J.* Using Correspondence Analysis to Combine Classifiers // Mach. Learn. — Hingham, MA, USA, 1999. — Vol. 36, no. 1/2. — P. 33–58.

96. A Novel Combining Classifier Method Based on Variational Inference / T. T. Nguyen [et al.] // Pattern Recogn. — New York, NY, USA, 2016. — Vol. 49, no. C. — P. 198–212.

97. *Петровский А. Б.* Методы групповой классификации многопризнаковых объектов (часть 1) // Искусственный интеллект и принятие решений. — 2009. — № 3. — С. 3—14.

98. *Петровский А. Б.* Методы групповой классификации многопризнаковых объектов (часть 2) // Искусственный интеллект и принятие решений. — 2009. — № 4. — С. 3—14.

99. Smart IDReader: Document Recognition in Video Stream / K. Bulatov [et al.] // 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Vol. 06. — 2017. — P. 39–44.

100. *Sourvanos N.*, *Tsatiris G.* Challenges in Input Preprocessing for Mobile OCR Applications: A Realistic Testing Scenario // 9th International Conference on Information, Intelligence, Systems and Applications (IISA). — 07/2018. — P. 1–5.

101. *Hartl A.*, *Arth C.*, *Schmalstieg D.* Real-time Detection and Recognition of Machine-Readable Zones with Mobile Devices // VISAPP 2015 - 10th International Conference on Computer Vision Theory and Applications; VISIGRAPP, Proceedings. Vol. 3. — 01/2015. — P. 79–87.

102. *Zilberstein S.* Using Anytime Algorithms in Intelligent Systems // AI Magazine. — 1996. — Sept. — Vol. 17. — P. 73–83.

103. An Anytime Algorithm for Camera-Based Character Recognition / T. Kobayashi [et al.] // 2013 12th International Conference on Document Analysis and Recognition. — 2013. — P. 1140–1144.

104. *А. Б. Б.* Задача наилучшего выбора / под ред. Т. Э. А. — Москва : Наука, 1984. — С. 196.

105. *Chow Y.*, *Robbins H.*, *Siegmund D.* Great expectations: the theory of optimal stopping. — Houghton Mifflin, 1971.

106. *Ferguson T. S.* Optimal Stopping and Applications. — 2008. — URL: https://www.math.ucla.edu/~tom/Stopping/Contents.html ; Accessed 13 November 2018.

107. *Tamaki M.* On the optimal stopping problems with monotone thresholds // Journal of Applied Probability. — 2015. — Vol. 52, no. 4. — P. 926–940.

108. *Mucci A. G.* On a Class of Secretary Problems // The Annals of Probability. — 1973. — Vol. 1, no. 3. — P. 417–427.

109. *Ferguson T. S.*, *Klass M. J.* House-hunting without second moments // Sequential Analysis: Design Methods and Applications. — 2010. — Vol. 29. — P. 236–244.

110. *Klass M. J.* Properties of Optimal Extended-Valued Stopping Rules for S_n/n^1 // The Annals of Probability. — 1973. — Vol. 1, no. 5. — P. 719–757.

111. *Ferguson T. S.*, *Hardwick J. P.* Stopping Rules for Proofreading // Journal of Applied Probability. — 1989. — Vol. 26, no. 2. — P. 304–313.

112. *Yang M. C. K.*, *Wackerly D. D.*, *Rosalsky A.* Optimal Stopping Rules in Proofreading // Journal of Applied Probability. — 1982. — Vol. 19, no. 3. — P. 723–729.

113. *Dalal S. R.*, *Mallows C. L.* When Should One Stop Testing Software? // Journal of the American Statistical Association. — 1988. — Vol. 83, no. 403. — P. 872–879.

114. Regularizing Neural Networks by Penalizing Confident Output Distributions / G. Pereyra [et al.] // CoRR. — 2017. — Vol. abs/1701.06548. — eprint: 1701.06548.

115. *Bulatov K.*, *Polevoy D.* Reducing Overconfidence in Neural Networks by Dynamic Variation of Recognizer Relevance // Proceedings of 29th European Conference on Modelling and Simulation (ECMS 2015) / ed. by V. M. Mladenov [et al.]. — 2015. — P. 488–491.

116. *Chow Y. S.*, *Robbins H.* A Martingale System Theorem and Applications // Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. Vol. 1. — Berkeley, Calif. : Univ. of Calif. Press, 1961. — P. 93–104.

117. *Sung Cheol Park*, *Min Kyu Park*, *Moon Gi Kang.* Super-resolution image reconstruction: a technical overview // IEEE Signal Processing Magazine. — 2003. — Vol. 20, no. 3. — P. 21–36.

118. A Survey: The Methods & Techniques of Super-Resolution Image Reconstruction / A. Semwal [et al.] // International Journal for Scientific Research & Development. — 2017. — Vol. 4, no. 12. — P. 243–249.

119. International standard ISO/IEC 14496-12: Information technology – Coding of audio-visual objects – Part 12: ISO base media file format. — ISO/IEC, 2005. — 94 p.

120. *Arlazarov V. L.*, *Loginov A. S.*, *Slavin O. A.* Characteristics of Optical Text Recognition Programs // Programming and Computer Software. — 2002. — May. — Vol. 28, no. 3. — P. 148–161.

121. *Fumera G.*, *Roli F.* Linear Combiners for Classifier Fusion: Some Theoretical and Experimental Results // Multiple Classifier Systems: 4th International Workshop, MCS 2003 Guildford, UK, June 11–13, 2003 Proceedings / ed. by T. Windeatt, F. Roli. — Berlin, Heidelberg : Springer Berlin Heidelberg, 2003. — P. 74–83.

122. *Schwenk H.*, *Gauvain J.-L.* Combining multiple speech recognizers using voting and language model information // IEEE International Conference on Speech and Language Processing. — 2000. — P. 915–918.

123. *Ye P.*, *Doermann D.* Document Image Quality Assessment: A Brief Survey // 2013 12th International Conference on Document Analysis and Recognition. — 2013. — P. 723–727.

124. *Николаев Д. П.*, *Полевой Д. В.*, *Чернов Т. С.* Метод автоматической оценки качества цветовой сегментации в задаче упаковки изображений печатных документов // Труды ИСА РАН. — 2013. — Т. 63, № 3. — С. 78—84.

125. Документооборот. Прикладные аспекты / под ред. В. Д. Арлазаров, Н. Е. Емельянов. — М. : Едиториал УРСС, 2005. — 184 с.

126. *Berend D.*, *Kontorovich A.* Consistency of Weighted Majority Votes // Proceedings of the 27th International Conference on Neural Information Processing Systems. — Montreal, Canada : MIT Press, 2014. — P. 3446–3454. — (NIPS'14).

127. *Džeroski S.*, *Ženko B.* Is Combining Classifiers with Stacking Better Than Selecting the Best One? // Mach. Learn. — Hingham, MA, USA, 2004. — Vol. 54, no. 3. — P. 255–273.

128. *Ilin D.*, *Krivtsov V.* Creating training datasets for OCR in mobile video stream // Proceedings of 29th European Conference on Modelling and Simulation (ECMS 2015) / ed. by V. M. Mladenov [et al.]. — 2015. — P. 516–520.

129. *Nitzan S.*, *Paroush J.* Collective Decision-Making and Jury Theorems // The Oxford Handbook of Law and Economics. Volume 1: Methodology and Concepts / под ред. F. Parisi. — 2017. — URL: http://www.oxfordhandbooks.com / view / 10 . 1093 / oxfordhb / 9780199684267 . 001 . 0001 / oxfordhb - 9780199684267-e-035 (дата обр. 28.06.2017).

130. *Chernov T.*, *Kolmakov S.*, *Nikolaev D.* An algorithm for detection and phase estimation of protective elements periodic lattice on document image // Pattern Recognition and Image Analysis. — 2017. — Vol. 27, no. 1. — P. 53–65.

131. Image quality assessment for video stream recognition systems / T. Chernov [и др.] // Proc SPIE. Т. 10696. — 2018. — С. 10696-10696—8.

132. *Журавлев Ю.* Распознавание. Классификация. Прогноз. Математические методы и их применение. Выпуск 2. — Москва : Наука, 1989. — 302 с.

133. *Sankoff D.*, *Kruskal J.* Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. — Center for the Study of Language, Information, 1999. — 408 p.

134. *Yujian L.*, *Bo L.* A Normalized Levenshtein Distance Metric // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2007. — Vol. 29, no. 6. — P. 1091–1095.

135. *Ding I. J.*, *Yen C. T.*, *Hsu Y. M.* Developments of Machine Learning Schemes for Dynamic Time-Wrapping-Based Speech Recognition. // Mathematical Problems in Engineering. — 2013. — P. 542680–1–10.

136. *Stuner B.*, *Chatelain C.*, *Paquet T.* LV-ROVER: Lexicon Verified Recognizer Output Voting Error Reduction // ArXiv e-prints. — 2017. — arXiv: 1707.07432.

137. *Cazenave T.* Overestimation for Multiple Sequence Alignment // CIBCB 2007: IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology. — 2007. — P. 159–164.

138. *Smith R.* An Overview of the Tesseract OCR Engine // Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02. Vol. 2. — IEEE Computer Society, 2007. — P. 629–633. — (ICDAR '07).

139. Method of determining the necessary number of observations for video stream documents recognition / V. V. Arlazarov [et al.] // Proc. SPIE. Vol. 10696. — 2018. — P. 10696-10696–6.

Список рисунков

Список таблиц