На правах рукописи

Шальнов Евгений Вадимович

Исследование и разработка методов сопровождения людей и частей их тела в видеопоследовательности

Специальность 05.13.11 —
«Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Автореферат
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2017

Работа выполнена в Московском Государственном Университете имени М.В. Ломоносова.

Научный руководитель: кандидат физико-математических наук, доцент
Конушин Антон Сергеевич

Официальные оппоненты: **Турлапов Вадим Евгеньевич,**
доктор технический наук, доцент,
Национальный исследовательский нижегородский университет им. Н.И.Лобачевского,
профессор

Бурцев Михаил Сергеевич,
кандидат физико-математических наук,
Московский физико-технический институт (государственный университет),
заведующий лабораторией нейронных систем и глубокого обучения

Ведущая организация: Государственный научный центр ФГУП «Государственный научно-исследовательский институт авиационных систем»

Защита состоится __ _____ 2018 г. в __ часов на заседании диссертационного совета Д 002.024.01 при Федеральное государственное учреждение «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук» по адресу: 125047, Москва, Миусская пл., 4.

С диссертацией можно ознакомиться в библиотеке ИПМ им. М.В. Келдыша РАН, `http://keldysh.ru`.

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, просьба направлять по адресу: 125047, Москва, Миусская пл., 4, ученому секретарю диссертационного совета Д 002.024.01.

Автореферат разослан __ _____ 2018 года.
Телефон для справок: +7 (499) 250-78-66.

Ученый секретарь
диссертационного совета
Д 002.024.01,
канд. физ.-мат. наук Бондарев Александр Евгеньевич

Общая характеристика работы

Актуальность темы.

В современном мире системы видеонаблюдения становятся важной частью инфраструктуры городов и предприятий. Под системой видеонаблюдения понимается комплекс программных и аппаратных средств получения и анализа видео для помощи в принятии решения человеком. В настоящее время в большинстве случаев системы видеонаблюдения используются для видеофиксации событий с целью последующего анализа и разбора человеком-оператором, например, после возникновения какой-либо внештатной ситуации. Ключевые вопросы, на которые необходимо ответить оператору: «кто присутствовал в видео?» и «какие события происходили?».

Текущий уровень развития алгоритмов компьютерного зрения позволяет автоматизировать получение ответов на эти вопросы для ряда важных практических сценариев. Достижения в решении задач выделения автомобилей на дороге и их идентификации по номерному знаку позволили создать систему автоматической фиксации нарушений правил дорожного движения, принуждающую водителей им следовать. В последние несколько лет были разработаны эффективные алгоритмы выделения лиц людей в видео и идентификации человека по изображению лица. На основе этих алгоритмов были созданы системы контроля доступа с идентификацией по лицу, автоматизации верификации личности по биометрическому паспорту на контрольно-пропускных пунктах или при оформлении кредитов и др.

Однако, потенциальные возможности видеонаблюдения существенно шире. До сих пор нерешенной остается задача идентификации в видео человека, чьё лицо скрыто, или его изображение имеет низкое разрешение. На рисунке 1 представлены примеры запечатленных противоправных действий. Хотя с помощью полученных данных и удается восстановить хронологию событий, но идентификация людей на кадрах во многих случаях потребует ручного труда, так как участники событий могут скрывать свои лица. В этой связи важным направленим развития является идентификация людей по особенностям комплекции и поведения, в частности походке. Также для идентификации человека важной является информация о траектории движения человека в поле зрения камеры или многокамерной системы. Она может позволить определить, откуда пришел, куда ушел интересующий человек, или найти момент времени, где его лицо еще не было скрыто маской или капюшоном. При этом задача сопровождения, то есть построения траектории движения, одного интересующего человека в видеопоследовательности сопряжена со значительными сложностями. Например, во многих случаях сложно выделить сопровождаемую цель в толпе из-за схожести комплекции или цвета одежды. Задача становится

Рис. 1 — Примеры кадров данных видеонаблюдения, на которых запечатлены противоправные действия. В первом ряду поджоги пианино и автомобиля. Во втором ряду взлом магазина и кража велосипеда.

еще сложнее, если искомый человек сознательно старается сбить со следа. В этой связи необходимо использовать сопровождение всех людей, присутствующих в видеопоследовательности. Даже если не удается выделить интересующего человека в толпе, этот подход позволяет определить траектории движения всех людей, находящихся рядом или похожих на интересующего, что существенно уменьшает сложность розыскной деятельности.

У задачи сопровождения всех людей в видеопоследовательности есть и другие применения. Её решение может упростить городское планирование за счет анализа количества и маршрутов движения людей и машин. Например, согласно отраслевому дорожному методическому документу ОДМ 218.6.003-2011 и ГОСТ 52289-2004, решение о необходимости

проектирования светофорного объекта принимается на основании результатов обследования транспортных и пешеходных потоков. Эти документы указывают плотность потока, при которой рекомендуется применять светофорное регулирование. Поэтому использование автоматических средств подсчета людей и машин позволит оперативно отслеживать изменение потоков движения и принимать решения в области городского планирования.

Однако, современные алгоритмы существенно уступают человеку в качестве сопровождения множества людей[1]. В связи с этим их использование для рашение практических задач очень ограничено. Другим существенным ограничением является высокая вычислительная сложность многих алгоритмов анализа видео, не допускающая их практическое применение на современном уровне развития техники. Широкая доступность видеокамер и развитие компьютерных сетей позволили создать системы видеонаблюдения, объединяющие более сотни тысяч камер. Однако даже алгоритмы первичного анализа такие, как обнаружение объектов интереса (людей, машин и др.), не позволяют обрабатывать больше нескольких видеопотоков на центральном процессоре или рассчитаны на дорогостоящие графические ускорители.

Одним из возможных решений проблемы высокой вычислительной сложности и низкого качества результатов обработки данных видеонаблюдения является использование информации о положении и свойствах используемой камеры, т.е. параметров её калибровки. Эта информация ограничивает возможные положения объектов интереса на кадрах, что может быть использовано как для уменьшения количества анализируемых регионов изображения, так и для обнаружений ложных срабатываний алгоритмов детектирования. К сожалению, существующие алгоритмы получения информации о камере либо требуют взаимодействие с пользователем и калибровочным шаблоном, либо могут быть примененты лишь для небольшого диапазона возможных положений камеры, что ограничивает их применимость.

Для развития систем видеонаблюдения необходимо разработать алгоритмы анализа, превосходящие существующие по точности и качеству. В своей работе я рассматриваю основной сценарий видеонаблюдения, включающих единственную неподвижную камеру. В рамках такой постановки стандартный подход к анализу данных видеонаблюдения, описанный в работе [1], заключается в решении следующих подзадач:

1. Калибровка камеры (построение отображения между мировой системой координат и системой координат изображения);
2. Обнаружение и сопровождение объектов интереса (например, людей) в видео;

[1] С результатами лучших современных алгоритмов сопровождения можно ознакомиться на странице соревнования MOTChallenge `https://motchallenge.net/`

3. Анализ поведения (подразумевает автоматическое определение типа поведения и выявление аномального поведения).

Целью данной работы является разработка методов повышения качества локализации, сопровождения и определения позы людей в видеопоследовательности, полученных статичной камерой, за счёт использования информации о калибровке камеры и движении людей в сцене.

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

1. Разработать и реализовать алгоритм определения положение и направления камеры в сцене по результатам обнаружения людей, допускающий определения угла наклона в пределах от 0 до $\frac{\pi}{2}$.
2. Разработать и реализовать алгоритм сопровождения каждого человека в видеопоследовательности, использующий информацию о калибровке камеры и регионах входа в сцену для повышения точности построения траекторий.
3. Разработать и реализовать алгоритм определения позы человека в видеопоследовательности, основанный на совместной модели положения и скорости движения суставов тела, позволяющий повысить точность решения задачи по сравнению предыдущим подходом.
4. На основе предложенных алгоритмов разработать программное средство для построения траекторий движения людей и их конечностей по видеопоследовательности, позволяющее решать поставленную задачу и допускающий использование различных алгоритмов локализации людей и визуального сопровождения путём замены отдельных модулей.

Основные положения, выносимые на защиту:

1. Предложен оригинальный метод определения положения и направления статичной камеры в сцене по результатам обнаружения людей, основанный на обучении отображения только на синтетических данных видеонаблюдения.
2. Для видеопоследовательностей, полученных статичной камерой, разработан алгоритм сопровождения людей, использующий положение и направление камеры для фильтрации ложноположительных срабатываний детектора.
3. Предложен алгоритм оценки позы человека в видеопоследовательности, учитывающий одновременно положение и скорость движения каждого сустава тела человека на кадре видеопоследовательности.
4. На основе предложенных алгоритмов разработан программный комплекс для автоматического сопровождения и опредеделния позы человека в видеопоследовательности и автоматизированное программное средство построения экспертной разметки позы человека на каждом кадре.

Научная новизна:

1. Впервые предложен алгоритм определения положения и направления статичной камеры в сцене по обнаружениям людей в видеопоследовательности, основанный на машинном обучении с возможностью настройки только на синтетических данных. Показано, что в отличие аналогов при анализе реальных данных видеонаблюдения точность предложенного алгоритма не уменьшается с увеличением угла наклона камеры от 0 до 90 градусов.
2. Впервые предложен алгоритм классификации обнаружений людей на изображении со статичной камеры на правдоподобные и недопустимые для данной сцены, основанный на машинном обучении с возможностью настройки только на синтетических данных. Показано, что применение предложенного алгоритма повышает скорость и среднюю точность обнаружения людей на изображении.
3. Впервые была предложены модель скелета человека, описывающая одновременно положение и движение каждого сустава человека в видеопоследовательности в виде линейной динамической системы. Показано, что ранее существовавшие модели являются частными случаями предложенной. На основе данной модели предложен новый алгоритм определения скелета (позы) человека в каждом кадре видео за сцет поиска локального оптимума целевого функционала. Предложенный алгоритм показал более высокую точность определения позы по сравнению с алгоритмами, основанными на предыдущих моделях.

Практическая значимость Одним из направлений развития видеонаблюдения является осуществление первых этапов обработки данных (в частности обнаружения объектов) ресурсами самой камеры. С учетом ограниченности вычислительных ресурсов, доступных на камере, предложенные в работе алгоритмы автоматической калибровки камеры и обнаружения людей в сцене имеют большую практическую значимость. Они позволяют расширить множество базовых алгоритмов обнаружения объектов, способных обработать изображение в при заданных ограничениях на время работы, то есть допускают использование более совершенных детекторов, которые, как правило, требуют больше вычислительных ресурсов.

Предложенный алгоритм детектирования людей, использующий информацию о калибровке, может применяться также и на PTZ-камерах, если количественная информация об изменении направления и фокусного расстояния поступает от приводов.

Предложенный алгоритм определения позы человека в видео допускает построение решения, соответствующего частичной экспертной разметке. На основе этой идеи было создано программное средство для построения

эталонной выборки позы человека в видео, состоящий из двух повторяющихся шагов:

- применение алгоритма поиска оптимальной позы человека в видео, соответствующего частичной экспертной разметке;
- расширение частичной экспертной разметки для исправления ошибок текущего решения.

Ценность предложенного средства заключается в существенном уменьшении ручного труда при разметке видеопоследовательностей. Такие размеченные данные являются ключевым фактором появления новых, более совершенных алгоритмов оценки позы человека в видео.

Предложенные алгоритмы были реализованы в виде программного средства (ПС). Разработанное ПС для построения траекторий движения людей и их конечностей в видео последовательности имеет модульную архитектуру, где каждый модуль решает отдельную задачу анализа входных данных. Замена модулей обеспечивает возможность повышения качества решения поставленных задач при использовании новых алгоритмов.

Апробация работы. Основные результаты работы докладывались на:

- семинаре им. М.Р. Шура-Бура под руководством М.М. Горбунова-Посадова;
- семинаре аспирантов кафедры АСВК и СКИ факультета ВМК МГУ под руководством Р.Л. Смелянского;
- международном семинаре МГУ-Huawei «избранные разделы обработки и анализа изображений» (CMC MSU-Huawei International Workshop "Selected topics in multimedia image processing and analysis"), Россия, Москва, 31 августа 2016;
- 5-м международном семинаре по анализу изображений (5th International Workshop on Image Mining. Theory and Applications), Берлин, Германия, 2015 год;
- 11-й международной конференции Распознавание образов и Анализ Изображений Россия, Самара, 2013 год;
- 26-й Международной конференции по компьютерной графике, обработке изображений и машинному зрению, системам визуализации и виртуального окружения GraphiCon 2016, Нижний Новгород, Россия, 19-23 сентября 2016 год;
- 25-й Международной конференции по компьютерной графике, обработке изображений и машинному зрению, системам визуализации и виртуального окружения GraphiCon 2015, Протвино, Россия, 22-25 сентября 2015 год;
- 24-й Международной конференции по компьютерной графике, обработке изображений и машинному зрению, системам визуализации и виртуального окружения GraphiCon 2014, Ростов-на-Дону, Россия, 30 сентября-3 октября 2014 год;

– летней школе Microsoft для аспирантов (Microsoft Research PhD Summer School), Англия, Кембридж, 2014.

Личный вклад. Личный вклад автора заключается в выполнении основного объёма теоретических и экспериментальных исследований, изложенных в диссертационной работе, включая разработку теоретических моделей, методик и разработку и реализацию алгоритмов, анализ и оформление результатов в виде публикаций и научных докладов.

В опубликованных работах А.С. Конушину принадлежит постановка задачи и обсуждение результатов её решения. Вклад В.С. Конушина состоит в построении обзора методов визуального сопровождения и обсуждении результатов.

Публикации. Основные результаты по теме диссертации изложены в 5 печатных изданиях, 4 из которых изданы в журналах, рекомендованных ВАК.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель, ставятся задачи работы, сформулированы научная новизна и практическая значимость представляемой работы.

Первая глава посвящена обзору научной литературы по изучаемой проблеме, описываются достоинства и недостатки существующих решений. Для каждой задачи выявлены недостатки существующих методов, обосновывающие актуальность данной диссертационной работы.

Вторая глава посвящена исследованию задачи определения положения и направления статичной камеры в сценарии видеонаблюдения.

Глава начинается с неформальной постановки рассматриваемой задачи. Задача калибровки камеры хорошо изучена, однако наиболее распространенные методы требуют взаимодействия с синтетическим объектом в сцене – калибровочным шаблоном, искусственно привносимым в сцену. Это ограничивает их применимость на практике. В данной главе рассматривается задача определения положения и направления (позы) камеры при отсутствии таких синтетических калибровочных шаблонов. В качестве калибровочных объектов предлагается использовать людей, запечатленных в видеопоследовательности.

В разделе 2.1 описывается модель наблюдаемых данных. Она состоит из моделей сцены, камеры и человека. Большинство сценариев видеонаблюдения содержат единственную плоскость земли, где могут находиться люди. Поэтому в данной работе рассматривется самая простая модель сцены, состоящая из горизонтальной плоскости земли и статичной камеры, расположенной на высоте h относительно неё. Предполагается, что свойства камеры описываются единственным параметром – её фокусным

расстоянием f. В качестве модели человека используется трехмерная модель, предложенная в работе [2].

Раздел 2.2 посвящен формальной постановке задачи определения положения и направления камеры. Разработанный алгоритм имеет два входных параметра: видеопоследовательность $\{I_t\}_t^T$ и фокусное расстояние f камеры, которой она получена. Выходом алгоритма является кортеж трех чисел: высоты h камеры над плоскостью земли и углов её наклона t и крена r. Также в разделе 2.2 описаны дополнительные ограничения, накладываемы на наблюдаемые данные. В частности описаны диапазоны значений положения и направления камеры и минимальный размер людей на изображении. Существующие подходы накладывают жёсткие ограничения на параметры направления камеры в сцене. Наиболее распространенными являются предположение об отсутствие крена и близости угла наклона к нулю. Поэтому одним из требований к разрабатываемому алгоритму является устойчивость к изменению угла наклона камеры, то есть независимость ошибки определения позы камеры от значения угла наклона камеры в диапазоне $\left[0, \frac{\pi}{2}\right]$.

В разделе 2.3 описан предложенный алгоритм определения положения и направления камеры в сцене. Он основан на использовании сверточной нейронной сети. В момент написания этой работы размер открытых выборок, содержащих информацию о положении людей и параметрах калибровки камеры, был недостаточным для обучения только по реальным данным видеонаблюдения. Поэтому обучение производилось на синтетической выборке видеонаблюдения. Раздел 2.3.1 отведён под описание способа построения синтетической выборки, соответствующей описанной ранее модели наблюдаемых данных. Используемая модель наблюдаемых данных является порождающей и позволяет синтезировать изображения с помощью стандартных методов растеризации. Для построения одного изображения необходимо задать:

- параметры камеры:
  - позу камеры в сцене (t, r, h);
  - фокусное расстояние f;
- положения людей на плоскости земли.

Значение параметров камеры определяет сцену. С помощью предложенного метода была построена синтетическая выборка данных видеонаблюдения, состоящая из 100374 различных сцен. Для каждой сцены было синтезировано не менее 200 человек, находящихся в ней.

В разделе 2.3.2 предлагается признаковое описания изображений. В качестве такого описания используется результаты детектирования голов людей на изображении. Это позволяет при построении синтетической выборки отказаться от моделирования несущественных деталей изображения, таких как текстура и поза людей (рис. 2).

В разделе 2.3.3 описываются способ построения прецедентов по результатам обнаружения голов людей и используемая архитектура нейронной сети. Построенная нейронная сеть предсказывает положение и направление камеры на основе 64 прямоугольников, ограничивающих изображения голов людей. Характер изменения размеров этих прямоугольников на изображении позволяет определить углы наклона и крена, а их средний размер – расстояние до плоскости земли. Выходами нейронной сети являются искомый вектор позы камеры и оценка точности предсказания каждого из параметров. В качестве функции потерь выступает отрицательный логарифм плотности нормального распределения в точке, соответствующей истинному значению позы камеры. При этом первый выход нейронной сети интерпретируется как математическое ожидание этого распределения, а второй – как стандартное отклонение каждого из предсказавыемых параметров.

Раздел 2.4 посвящен обучению и экспериментальной оценки предложенного алгоритма. Обучение нейронной сети производилось на построенной синтетической выборке. Для устойчивости результатов к ошибкам обнаружения головы человека в синтетические данные были добавлены равномерно распределенные ложноположительные срабатывания детектора.

Экспериментальная оценка предложенного алгоритма проводилась на нескольких выборках. При тестировании на тестовой части синтетической выборки предложенный алгоритм показал отсутствие зависимости ошибки оценки позы камеры от истинного значения угла её наклона. В то же время показано, что при использовании ранее существовавшего подхода увеличение угла наклона камеры приводи к увеличению ошибки оценки угла наклона и высоты камеры. На выборке TownCentre производилась оценка точности работы алгоритма при наличии и отсутствии ложных обнаружений детектора. В обоих случаях предложенный алгоритм предсказал углы наклона и крена камеры с погрешностью не превосходящей 1.5 градусов, а высоту камеры с точностью до 80см. Также эксперименты показали, что с увеличением количества ложных обнаружений ошибка предсказания увеличивается, но остается в пределах трёх значений оценки точности, предсказанной сетью. При тестировании на более сложной выборке PETS2006 нейронная сеть не смогла корректно определить положение камеры для некоторых сцен. Это связано с увеличением количества ложных срабатываний детектора на данной выборке по сравнению с обучающей. Также в этих сценах детектор находил людей на двух плоскостях – первом и втором этажах, что нарушает одно из базовых предположений о наблюдаемой сцене.

Третья глава посвящена исследованию задачи обнаружения людей на изображении в сценарии видеонаблюдения с известными параметрами калибровки статичной камеры.

(а) (б)

Рис. 2 — Пример изображения реальных (а) и синтетических данных (б). Красным выделены результаты работы алгоритма обнаружения изображения головы человека.

Глава начинается с постановки рассматриваемой задачи. Задача обнаружения объектов интереса на изображения — базовая задача компьютерного зрения. Классическим подходом к её решению является метод скользящего окна, основанный на классификации различных регионов изображения на несколько классов (в частном случае на два — «объект интереса», «другое»).

В общем случае алгоритмы обнаружения не могут полагаться на априорное знание о положении и размерах объектов интереса на изображении и вынуждены перебирать огромное количество комбинаций этих параметров, многие из которых соответствуют неправдоподобным параметрам размера человека в рассматриваемой сцене. Исключение таких регионов из обрабатываемой части изображения позволяет 1) повысить скорость обработки изображения и 2) уменьшить вероятность ложного срабатывания.

В разделе 3.1 описывается предложенный алгоритм обнаружения людей на изображении, использующий известные значения параметров калибровки камеры. Формально, задача состоит в построении алгоритма, входом которого являются входное изображение I, фокусное расстояние камеры f, вычисленное в пикселях, и параметры положения и направления камеры l_c, а выходом является множество прямоугольников, ограничивающих изображения объектов интереса (людей или частей их тела). Предложенный алгоритм является суперпозицией $g(A_b, A_f)$ базового алгоритма A_b обнаружения людей на изображении, не использующего

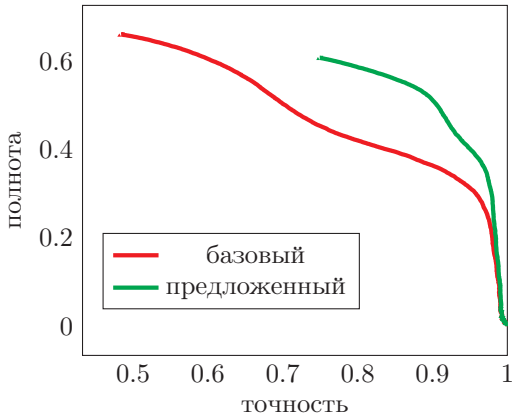

Рис. 3 — График зависимости полноты/точности обнаружения изображений голов людей. Оценка проводилась по выборке TownCentre [3].

информацию о калибровке камеры, и классификатора A_f результатов обнаружения на правдоподобные и недопустимые для данной сцены. Таким образом, разработка алгоритма обнаружения сводится к задаче построения классификатора A_f. Формально, входом предложенного классификатора является ограничивающий прямоугольник обнаруженного объекта o и вектор параметров калибровки камеры c, объединяющий фокусное расстояние f и параметры положения и направления камеры l_c.

Построение классификатора A_f происходит с помощью машинного обучения на множестве прецедентов. В качестве классификатора A_f используется полносвязная нейронная сеть, обученная на прецедентах синтетической выборки, описанной в предыдущей главе. Поскольку построенная выборка содержит только верные обнаружения объектов в сцене, то задачу построения классификатора является задачей поиска аномалий в данных. Однако путём построения прецедентов отрицательного класса она была сформулирована в виде обучения бинарного классификатора. Для построения отрицательной выборки в работе предлагается объединение двух стратегий: 1) использование обнаружений, характерных для других сцен обучающей выборки (параметров калибровки камеры); 2) использование регионов изображения с произвольным положением и размером. Первая стратегия позволяет классификатору выявлять обнаружения, характерные для каждой сцены. Вторая позволяет классификатору отличать верные обнаружения от ложных срабатываний базового алгоритма обнаружения A_b. При построении отрицательных примеров обучающей выборки эти стратегии используются в отношении $1 : 9$.

В разделе 3.2 описываются процесс обучения предложенного классификатора и его качество на синтетической и реальной выборках. При

(а) (б)

Рис. 4 — Маски обрабатываемых регионов, соответствующих размеру головы человека 28×28 пикселей. (а) маска, предсказанная классификатором, (б) маска, используемая для ускорения алгоритма обнаружения.

тестировании в качестве базового алгоритма A_b также используется оптимизированная реализация алгоритма поиска голов людей на изображении [4]. Используемая в качестве классификатора нейронная сеть состоит из 5 слоёв. Её гиперпараметры (количество слоёв и их размер) настраивались по синтетической валидационной выборке. Тестирование алгоритма проводилось на синтетической тестовой выборке и реальных данных видеонаблюдения.

Предложенный метод настраивает классификацию для всевозможных значениях высоты камеры в диапазоне от 0 до 20 метров и значении угла её наклона от 0 до $\frac{\pi}{2}$. Эти ограничения связаны с диапазоном значений, в котором находятся параметры при построении синтетической выборки. Результаты на синтетической тестовой выборке показали площадь под ROC кривой равную 0.926. При этом полнота обнаружения положительных примеров тестовой выборки при стандартном пороге 0.5 составляет 0.988. При применении к тестовой части синтетической выборки построенный классификатор показал устойчивость к значениям параметров наклона и крена камеры. Множества сцен обучающей, валидационной и тестовых выборок не пересекаются. Таким образом, результаты тестирования показывают, что предложенный классификатор не переобучается. Тестирование на реальных данных показало, что построенный классификатор позволяет существенно увеличить точность обнаружения людей без существенного падения полноты (рис. 3).

Также в работе рассматривается задача уменьшения вычислительной сложности обработки изображения базовым алгоритмом обнаружения A_b при известных параметрах калибровки камеры. Предложенный классификатор для каждого размера головы определяет регионы, где изображение голов людей имеют «правдоподобные» размеры (рис. 4 (а)). Обработка только этих регионов позволяет повысить скорость и точность обработки данных. Из-за особенностей реализации алгоритмов обнаружения производить обработку изображения по произвольной маске неэффективно.

Поэтому в предложенной реализации обработка изображения каждого масштаба производится в наименьшем ограничивающем прямоугольнике, содержащем обнаруженную область (рис. 4 (б)). Экспериментальная оценка показала, что предложенный классификатор в среднем распознает только 21.19% всех просматриваемых окон изображения как «правдоподобные». При обработке изображения по прямоугольной маске количество обрабатываемых окон в среднем увеличивается до 38.76%. Существенное увеличение количества обрабатываемых окон характерно для сцен, где крен камеры значительно отличается от нуля. Однако такой сценарий редко встречается в практических сценариях видеонаблюдения. Например, при использовании предложенного классификатора для сцены выборки TownCentre [3] «правдоподобными» являются только 21.44% окон, рассматриваемых базовым алгоритмом A_b, а предложенный метод обрабатывает 24.03% окон. Эмпирически показано, что обработка изображения в соответствии с построенной маской увеличивает производительность базового алгоритма на тестовой выборке в среднем на 41%.

Четвертая глава посвящена исследованию задачи сопровождения людей в видеопоследовательности. Под сопровождением понимается построение траектории движения каждого человека, присутствующего в видео, а траектория описывается последовательностью ограничивающих прямоугольников человека, на каждом кадре, где он присутствовал. Наиболее перспективным считается подход сопровождения через обнаружение, разбивающий задачу на два этапа: 1) поиск людей в видео и 2) объединение результатов обнаружения в траектории.

В разделе 4.1 подробно описан существующий подход решения задачи сопровождения [3] и обоснован выбор его в качестве базового алгоритма. Данный метод формулирует задачу построения траекторий в терминах минимизации целевого функционала, задающего ненормированное распределение на множестве наблюдаемых данных. Также в разделе описаны недостатки базового подхода, устранение который позволяет повысить точность построения траекторий.

В разделе 4.2 описаны предложенный алгоритм и его отличие от базового. Ключевым элементом целевого функционала является потенциал $\phi_f(d^{t_2}|d^{t_1})$, описывающий штраф за нахождения в одной траектории результатов обнаружения d^{t_1} и d^{t_2}, соответствующих кадрам I_{t_1} и I_{t_2}, $t_1 < t_2$. В базовой работе оценивается, насколько вероятно, что человек придёт в область d^{t_2} в момент времени t_2, учитывая траекторию его движения в локальной окрестности предшествующего кадра I_{t_1} (рис. 5 (а)). В качестве модификации предлагается использовать также фактор $\phi_b(d^{t_1}|d^{t_2})$, описывающий сопровождение в обратном направлении. Его можно интерпретировать, как проверку того, что человек, оказавшийся в области d^{t_2}

(а) (б)

Рис. 5 — Визуализация способа построения фактора $\phi(d^{t_2}|d^{t_1})$ сходства обнаружений. Зелеными прямоугольниками отмечены результаты работы детектора головы человека на кадрах I_{t_1} и I_{t_2}, зеленой кривой обозначена траекторий движения человека в локальной окрестности его обнаружения, красные прямоугольники соответствуют построенным гипотезам положения его головы. Фактор $\phi_f(d^{t_2}|d^{t_1})$ описывает сходство результатов обнаружения при сопровождении в прямом направлении (а), а фактор $\phi_b(d^{t_1}|d^{t_2})$ — при сопровождении в обратном (б).

Рис. 6 — Регион входа в сцену, полученный при сопровождении в течении 3-х минут.

в момент времени t_2, был в области d^{t_1} на кадре I_{t_1}(рис. 5 (б)). В предложенном алгоритме фактор сходства результатов обнаружения d^{t_1} и d^{t_2} вычисляется как среднее арифметическое факторов ϕ_f и ϕ_b.

Второй модификацией предложенного алгоритма является использование информации о регионах входа в наблюдаемую сцену. Эта модификация учитывает предположение, что траектории должны начинаться в областях входа, то есть нацелена на уменьшение фрагментации (количества разрывов) траекторий движения людей. В оптимизируемый целевой функционали добавлен фактор, отдающий приоритет траекториям, начало

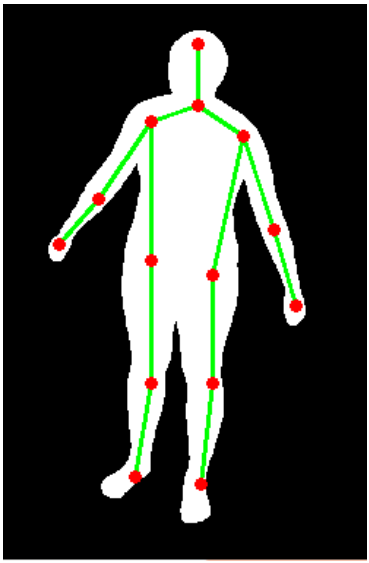

Рис. 7 — Визуализация понятия позы человека на изображении. Красными точками отмечены суставы p_i^t, описывающие позу человека, отрезками обозначены семантические части.

которых находится вблизи региона входа: $\psi(d^0) = \rho(d^0, S)$, где ρ определяет l_2 расстояние между центром ограничивающего прямоугольника обнаружения и регионом входа S. В качестве региона входа используется граница выпуклой оболочки наиболее надёжных построенных траекторий. На рисунке 6 представлен регион входа в сцену видеопоследовательности TownCenter [3], полученный при сопровождении в течении 3-х минут.

В пункте 4.3 представлены результат сравнения предложенного алгоритма с базовым. Сравнение показало повышение точности построения траектории людей на 0.09 по метрике MOTA при использовании предложенных модификаций. Также результаты содержат анализ влияния различных этапов работы алгоритма на точность построения траекторий. Анализ показывает, что наибольшего повышения точности (до 0.287) по метрике MOTA можно получить за счёт дальнейшего развития целевого функционала, определяющего способ объединения результатов обнаружения людей в траектории.

В **пятой главе** приведено описание предложенной модели положения и движения суставов человека в видео. Также предложены алгоритмы поиска локального оптимума предложенной модели. Положение множества суставов тела человека на изображении I_t образуют его позу P^t, т. е. $P^t = \{p_i^t\}_{i=1}^K$, где $p_i^t \in \mathcal{R}^2$ — положение i-ого сустава тела человека на кадре I^t (рис. 7). Расширение этого понятия на случай видеопоследовательности $\{I_t\}_{t=1}^T$ подразумевает определение позы человека на каждом кадре:

$P = \{P^t\}_{t=1}^T$. В практических реализациях поза человека считается дискретной величиной, где возможное положение суставов ограничено узлами регулярной сетки на кадре.

В разделе 5.1 описывается новая математическая модель положения и движения суставов человека в видео. Предложенная модель определяется целевым функционалом $E(P, \Theta)$, минимум которого соответствует оптимальной позе человека в видео. Минимизация проводится по значению позы человека в видео P и скрытым параметрам модели Θ. В существующих работах целевой функционал представляют в виде суммы факторов двух типов: модели позы человека на изображении $E_I(P_t, \Theta)$ и модели изменения позы между кадрами $E_T(P, \Theta)$.

В разделе 5.1.1 обосновывается выбор модели из набора частей в качестве модели $E_I(P_t, \Theta)$ позы человека на изображении. Важно отметить, что модель из набора частей не является конкретной реализацией алгоритма определения позы человека, а описывает целый класс алгоритмов, объединённых общими свойствами. Дальнейшее рассмотрение задачи определения позы человека в видео не зависит от выбора конкретного алгоритма позы человека на изображении. Для проведения экспериментов была выбрана реализация модели [5], поскольку она использовалась в аналогичных работах, а следовательно позволяет оценить вклад в изменение точности при использовании предложенной модели движения суставов. В качестве скрытого параметра базовой модели выступает дискретный параметр $s^t \in \Theta$ размера позы человека на изображении.

В разделе 5.1.2 описана предложенная модель изменения позы между кадрами. В существующих работах в качестве функции изменения позы использовалось L_2 расстояние между соответствующими суставами позы на соседних кадрах. В данной работе предложено расширение этой модели, учитывающее характеристики движения суставов позы. Для этого скрытые параметры Θ модели расширены непрерывными параметрами скорости движения суставов на каждом кадре. Предложенный целевой функционал использует L_2 расстояние для моделирования отклонения изменения позы от линейной модели движения. Предложенная модель изменения позы имеет вид:

$$E_T(P, \Theta) = \sum_{i=1}^{K} \left(\psi_i^0(v_i^1) + \sum_{t=1}^{T-1} \psi_i^t(h_i^{t+1}, h_i^t, \Theta) \right) + \sum_{t=1}^{T-1} \eta^t(s^{t+1}, s^t),$$

где $\psi_i^0(v_i^1)$ определяет априорное предпочтение на скорость движения сустава на первом кадре, $\psi_i^t(h_i^{t+1}, h_i^t, \Theta)$ задаёт модель изменения состояния сустава с течением времени с учетом линейной модели движения, $\eta^t(s^{t+1}, s^t)$ — модель изменения размера позы.

В разделе 5.1.3 описаны два частных случая предложенной модели позы. Показано, что предложенная модель позы является обобщением

использованной ранее модели [5] и содержит её в качестве частного случая. Другим частным случаем является независимое определение позы человека на кадрах видеопоследовательности без использования модели изменения позы.

Разделе 5.2 посвящён описанию метода поиска локального минимума предложенного целевого функционала. В разделе 5.2.1 доказаны два важных свойства предложенной модели. Показано, что при известной позе человека в видео, задача определения оптимальных параметров скорости имеет точное решение, сложность нахождения которого линейно зависит от длины T видеопоследовательности, количества K суставов позы человека и количества M возможны положений каждого из них на изображении. Второе свойство предложенной модели связано с условной моделью $E(P^t, s^t | P^{\setminus t}, \Theta_{\setminus s^t})$ позы P^t и размера s^t человека на кадре I_t, обусловленной известными значениями остальных параметров модели. Доказывается возможность поиска глобального минимума рассматриваемой условной модели с вычислительной сложностью $\mathcal{O}(MK)$, где M — количество возможных положений на изображении каждого из K суставов.

В разделе 5.2.2 описан новый детерминированный алгоритм поиска локального минимума предложенного целевого функционала. Предложенный алгоритм основан на блочной оптимизации по множеству дискретных и непрерывных переменных модели и использует описанные выше свойства модели. В подпункте 5.2.3 предложен новый стохастический алгоритм уточнения результата, полученного детерменированным алгоритмом. Предложенный стохастический алгоритм основан на построении выборки из распределения, соответствующего целевому функционалу $E(P) = \min_\Theta E(P, \Theta)$.

Раздел 5.3 содержит сравнение результатов применения предложенной модели с базовой [5]. Первые результаты показаны на данных, используемых в [5]. Эти данные являеются наиболее сложными для предложеного алгоритма. Они содержат движение камеры и изменение масштаба съемки, поэтому предположение о линейности движения суставов на изображении не может быть выполнено. Однако результаты показали, что предложенная модель позволяет повысить точность локализации суставов на всех примерах открытой выборки, за исключением одного. Следующий эксперимент проводился на выборке TownCentre, полученной статичной камерой видеонаблюдения. Этот сценарий наиболее близко соответствует рассматриваемым предположениями о данных. На выборке TownCentre предложенный алгоритм показал значительный прирост точности (+8.7%) определения позы человека.

Шестая глава посвящена программной реализации разработанных алгоритмов и построение программных средств на их основе.

В разделе 6.1 описывается программное средство для сопровождения людей и определения их поз в видео. Данное программное средство построенно на основе всех алгоритмов, предложенных в предыдущих главах. С его помощью осуществляется локализация камеры, обнаружение людей на изображении, их сопровождение и определение позы.

Данное программное средство применяется в компании ООО «Технологии видеоанализа» для подсчета людей, прошедших сигнальную линию. За счет использования информации о калибровке камеры, оно позволяет повысить частоту применения детектора людей к кадрам входного видео. Благодаря этому, удается повысить точность подсчета людей в условиях обработки видеопотока.

В разделе 6.2 описывается полуавтоматическое средство разметки позы человека в видеопоследовательности, основанное на предложенном в главе 5 алгоритме. Построенное средство позволяет пользователю исправлять ошибки локализации суставов тела человека и повторять поиск наилучшей позы с учетом этих новых ограничений. Показано, что пользовательский ввод, ограничивающий область возможных положений сустава тела человека, не меняет структуру целевого функционала и сложность оптимизации. За счет уточнения положения лишь небольшого множества суставов на разреженном множестве кадров удается добиться сущетственного ускорения построения экспртной разметки позы людей в видеопоследовательности.

Данное программное средство было выполнено в рамках работы по проекту РФФИ 16-29-09612 офи_м «Исследование и разработка методов биометрической идентификации человека по походке, жестам и комплекции в данных видеонаблюдения». С его помощью происходит построение эталонной коллекции данных видеонаблюдения с известной позой человека на каждом кадре. Генерируемая коллекция используется для решения задачи идентификации человека по походке.

В **заключении** приведены основные результаты диссертационной работы и рассмотрены возможные варианты их применения.

1. Предложен оригинальный метод определения положения и направления статичной камеры в сцене по результатам обнаружения людей.
2. Для видеопоследовательностей, полученных статичной камерой, разработан алгоритм сопровождения людей, использующий положение и направление камеры для фильтрации ложноположительных срабатываний детектора.
3. Предложен алгоритм оценки позы человека в видеопоследовательности, учитывающий одновременно положение и скорость движения каждого сустава тела человека на кадре видеопоследовательности.

4. На основе предложенных алгоритмов разработан программный комплекс для автоматического сопровождения и опредеделния позы человека в видеопоследовательности и автоматизированное программное средство построения экспертной разметки позы человека на каждом кадре.

Дальнейшее развитие предложенных алгоритмов возможно по следующим направлениям:

- Оценка как положения и направления, так и фокусного расстояния камеры за счет использования результатов определения всей позы человека;
- Использование алгоритмов реидентификации человека по изображению, для надежного сопоставления людей в траектории;
- Добавление зависимости от входной видеопоследовательности в факторы гладкости изменения положения суствов тела человека.

Публикации автора по теме диссертации

1. *Шальнов Е. В.*, *Конушин А. С.* Использование геометрии сцены для увеличения точности детекторов // Программные продукты и системы. — 2017. — Т. 30, № 1. — С. 106—111.
2. *Shalnov E.*, *Konushin V.*, *Konushin A.* An improvement on an MCMC-based video tracking algorithm // Pattern Recognition and Image Analysis. — United States, 2015. — Vol. 25. — P. 532—540.
3. *Shalnov E.*, *Konushin A.* Human Pose Estimation in Video via MCMC Sampling // Proceedings of the 5th International Workshop on Image Mining. Theory and Applications. — 2015. — P. 71—79.
4. *Shalnov E.*, *Konushin A.* Convolutional Neural Network for Camera Pose Estimation from Object Detections. // International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences. — 2017. — Т. 42.
5. *Shalnov E. V.*, *Konushin V. S.*, *Konushin A. S.* Improvement of MCMC-based video tracking algorithm // 11th International Conference on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-11-2013). Samara, September 23-28, 2013. Conference Proceedings. Vol. 2. — IPSI RAS Samara, 2013. — P. 727—730.

Список литературы

1. *Wang X.* Intelligent multi-camera video surveillance: A review // Pattern recognition letters. — 2013. — Т. 34, № 1. — С. 3—19.
2. Building Statistical Shape Spaces for 3D Human Modeling / L. Pishchulin [и др.] // arXiv. — 2015. — Март.

3. *Benfold B.*, *Reid I.* Stable multi-target tracking in real-time surveillance video // Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. — IEEE. 2011. — С. 3457—3464.

4. *Prisacariu V.*, *Reid I.* fastHOG - a real-time GPU implementation of HOG: тех. отч. / Department of Engineering Science, Oxford University. — 2009. — № 2310/09.

5. *Park D.*, *Ramanan D.* N-best maximal decoders for part models // 2011 International Conference on Computer Vision. — IEEE. 2011. — С. 2627—2634.

Шальнов Евгений Вадимович

Исследование и разработка методов сопровождения людей и частей их тела в видеопоследовательности

Автореф. дис. на соискание ученой степени канд. физ.-мат. наук

Подписано в печать ___.___._____. Заказ № _____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Типография _____