Московский Государственный Университет имени М.В. Ломоносова

На правах рукописи

УДК 004.932.4

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»

Диссертация на соискание учёной степени
кандидата физико-математических наук

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

Москва — 2017

Оглавление

Стр.

Стр.

Стр.

Введение

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 0.1 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

Рисунок 0.1 — Примеры кадров данных видеонаблюдения, на которых запечатлены противоправные действия. В первом ряду поджоги пианино и автомобиля. Во втором ряду взлом магазина и кража велосипеда.

направлением 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 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

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

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 применены 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;
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» (СМС 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.

Объем и структура работы. Диссертация состоит из введения, шести глав и заключения. Полный объём диссертации составляет 115 страниц, включая 22 рисунка и 7 таблиц. Список литературы содержит 65 наименований.

Глава 1. Обзор литературы

1.1 Определение положения и направления камеры

В компьютерном зрении задача определения положения и направления камеры в сцене, называемых также позой камеры, исследуется давно [2—9]. Её решением является метод построения отображения мировой системы координат в систему координат, связанную с камерой. Входными данными для построения этого отображения являются кадры, полученные камерой.

В работе [8] представлен подход, извлекающий информацию о PTZ-камере при её движении. Авторы использовали сопровождение ключевых точек при повороте камеры и изменении масштаба. Это позволило оценить фокусное расстояние камеры и направление осей мировой системы координат. В то же время большое количество камер видеонаблюдения являются статичными, то есть не изменяют своего положения и направления в сцене с течением времению.

В своей работе я рассматриваю сценарий неподвижной камеры. Для него можно выделить два подхода к решению задачи оценки позы камеры. Алгоритмы первого подхода анализируют особенности распределения направлений прямых на изображении сцены для восстановления орентации мировой системы координат. Методы, отнесенные ко второму подходу, используют наблюдаемое распределение размеров известных объектов, таких как люди или автомобили, в разных частях изображения.

1.1.1 Анализ распределения направлений прямых на изображении

Методы первого подхода предполагают, что сцена представляет, так называемый, «Манхэттенский мир». Это определение описывает сцены, созданные

человеком, где преобладают три ортогональных направления прямых: два горизонтальных и одно вертикальное. Обычно в качестве направления осей мировой системы координат в работах предлагается выбирать направления этих прямых. Перспективные преобразования приводят к тому, что изображения этих прямых пересекаются в трех соответствующих точках схода. Так точки схода горизонтальных прямых лежат на линии горизонта, а точка пересечения вертикальных прямых образует зенит или надир. В условиях предположения «Манхэттенского мира» в описанных трех точках схода пересекается наибольшее количество прямых изображения. Работы первого подхода направлены на локализацию этих точке схода на изображении. Для краткости в дальнейшем точки схода, соответствующие ортогональным прямым сцены, я буду называть ортогональными. В работе [2] представлено соотношение между положением трех ортогональных точек схода и фокусным расстоянием камеры. Оно послужило основой для последующих алгоритмов определения позы камеры. В работе [3] предлагается извлекать ортогональные прямые из изображения объектов, таких как здания. Однако предложенный метод не может быть применен в сценах, где такие структуры отсутствуют, или не все необходимые точки схода могут быть найдены. Поэтому в работе [9] предлагается использовать видимое направление движения автомобилей по автостраде для извлечения горизонтальных прямых на изображении. Авторы [4; 5] используют направление движения людей и ориентацию их изображения для поиска линии горизонта и вертикальной точки схода. В рамках такого подхода люди в сцене описываются вертикальными отрезками. Точность такой модели существенно понижается, когда направление съемки камеры отличается от горизонтального. Поэтому в своей работе я не использовал информацию о ориентации изображения человека для оценки позы камеры.

1.1.2 Анализ размеров объектов на изображении

Алгоритмы второго подхода анализируют распределение размеров известных объектов на изображении сцены. Классическим предположением этих методов является наличие единственной плоскости земли, на которой располагаются все объекты. Самый известный алгоритм этого подхода был предложен в работе [10]. Авторы построили вероятностную графическую модель, описывающую зависимость между положением камеры и размерами людей и машин в сцене. В работе [7] построена функция зависимости позы камеры от размера человека в центре изображения и положения горизонта. Предложенные алгоритмы имеет ряд существенных ограничений. Авторы предполагают, что границы прямоугольников, ограничивающих изображение человека, совпадают с положением верхней и нижней точкой человека в сцене. Это условие выполняется, если направление съемки камеры близко к горизонтальному, и становится абсолютно неверным, когда камера направлено вертикально вниз. В описанных работах не учитывается возможность появления ложных обнаружений объектов и их влияние на результаты оценки позы камеры. Также авторам приходится ограничиваться случаем отсутствия крена камеры для построения аналитических формул отображения размеров объектов в положение камеры.

В своей работе я предлагаю алгоритм оценки позы камеры на основе анализа размеров голов людей в сцене. В отличие от предыдущих методов предложенных алгоритм оценивает наклон камеры в диапазоне $\left[0, \frac{\pi}{2}\right]$ и крен камеры в диапазоне $\left[-\frac{\pi}{6}, \frac{\pi}{6}\right]$. В своей работе я допускаю наличие ложных срабатываний детектора, и предложенный метод адаптируется к ним, оценивая ошибку в предсказании позы камеры по наблюдаемым данным.

1.2 Локализация объектов

Задача построения детектора объектов на изображении всегда интересовала исследователей в области компьютерного зрения. Современные алгоритмы обнаружения объектов работают по принципу скользящего окна, который разбивает цикл обработки на два этапа: 1) построение множества гипотез положения объекта на изображени, называемых окнами, и 2) классификация изображения внутри окна. Обычно от разрабатываемых алгоритмов требуется как можно более высокая скорость обработки данных и минимально возможное количество ложных срабатываний. Эти ограничания противоречат друг другу. И часто повышение точности детектирования приводит также к повышению его вычислительной сложности. Для практического применения в видеонаблюдении скорость обработки данных является ключевым параметром. Поэтому большое количество исследований посвящено способам уменьшения вычислительной сложности детектирования объектов при сохранении их качества. Можно выделить два основных направления работы в этой области: построение быстрого классификатора окон и уменьшение их количества.

1.2.1 Построение быстрых классификаторов

Исторически первые работы по ускорению детектирования посвящены ускорению применяемого классификатора. Авторы [11] предложили использовать каскад простых классификаторов для детектирования лиц на изображениях. Первые этапы каскада отбрасывают большое количество «простых» для классификации окон, не содержащих лиц, уменьшая общее время классификации. Предложенная идея оказалась настолько эффективной, что каскадные

детекторы стали применяться даже в цифровых фотоаппаратах. Одним из важных недостатков такого подхода является отсутствие возможности изменять соотношение точность/полнота для уже построенного классификатора. В работе [12] преодолевают это ограничение, изменив структуру каскада. Авторы предлагают так называемый «мягкий» каскад, в котором разделены этапы построения простых классификаторов каскада и выбор границы для разделения положительных и отрицательных примеров. Это позволяет настраивать полученный каскад под требования к точности без переобучения классификаторов. В работе [13] добиваются ускорения классификатора за счет вычисления признаков лишь на разреженной пирамиде изображений. На промежуточных слоях авторы предлагают восстанавливать признаки с помощью интерполяции.

В своей работе я предлагаю алгоритм понижения вычислительной сложности детектора, который не зависит от типа используемого классификатора окон, поэтому его можно использовать совместно с быстрыми классификаторами.

1.2.2 Уменьшение количества окон

Другое направление по ускорению обнаружения объектов посвящено уменьшению количества рассматриваемых окон. Авторы работы [14] используют корреляцию откликов классификатора в соседних окнах для выделения регионов изображения, где могут находиться объекты. Для этого на первых этапах обработки производится классификация лишь разреженного множества окон. Позднее детально анализируются только области, рядом с которыми были получены положительные отклики классификатора. В связи с существенным успехами нейросетевых алгоритмов классификации изображений [15—18] сверточные нейронные сети стали применять и для задачи обнаружения объектов на изображении. Обычно нейросетевые классификаторы требуют больших

вычислительных ресурсов. Поэтому в работе [19] предлагается применить нейросетевой классификатор лишь для небольшого подмножества окон, выбранных на изображении. В работах [20; 21] развивается предыдущая идея и предлагается разбить нейросетевой классификатор окон на этапы выбора интересующих регионов окна и уточнения положения объекта. Это позволило увеличить размеры окон и уменьшить их количество.

Предложенный мной в диссертационной раоботе алгоритм может быть интегрирован с любым из предложенных методов уменьшения количества обрабатываемых окон. Он дает априорную оценку расположения областей, где могут находиться объекты интереса, то есть ограничивает обрабатываемую область изображения еще до применения детектора.

1.3 Сопровождение объектов

Сопровождение объектов заключается в построении траекторий их движения в видеопоследовательности. В стандартной постановке рассматривается движение в системе координат изображения, а не наблюдаемой сцены. Существует два подхода к решению задачи: визуальное сопровождение и сопровождение через обнаружение. Первый подход применяется для построения траекторий объектов в видео, когда тип отслеживаемого объекта не известен. Сопровождение через обнаружение используется только для тех классов объектов, которые могут быть локализованы помощью детектора.

1.3.1 Визуальное сопровождение

Визуальное сопровождение применяется для локализации объектов, когда его тип не известен. Алгоритмы, относящиеся к этому подходу, одинаково

подходят для построения траекторий движения как изображений лиц или машин, так и для отслеживания перемещения в видео более абстрактных регионов изображения. Поэтму для построения внутреннего представления (шаблона) сопровождаемого объекта таким алгоритмам необходимо указать его положение на первом кадре. На последующих кадрах происходит поиск регионов изображения, наиболее похожих (по некоторому критерию) на указанный.

Алгоритмы визуального сопровождения различаются используемыми представлениями объектов и способами определения положения объекта на последующих кадрах. Одним из наиболее простых методов поиска на последующих кадрах является кросс-корреляция шаблонов [22]. При этом представлением объекта является его изображение на первом кадре. Этот подход позволяет добиться высокой скорости работы, однако он неустойчив к изменениям сопровождаемого объекта. В частности из-за движения разных частей тела друг относительно друга он плохо подходит для построения траектории фигуры идущего человека. Поэтому сопровождение объектов на основе кросс-корреляции шаблонов используется либо для отслеживания частей тела человека, либо для обработки таких коротких фрагментов видео, что существенных изменений изображения объекта не происходит.

Широкое распространение получил алгоритм сопровождения, названный фильтром частиц [23]. Его особенностью является метод поиска возможных положений объекта на следующем кадре. В отличие от предыдущего метода алгоритм [23] описывает положение объекта на кадре не одной точкой, а случайной величиной, определенной на двумерном пространстве. Плотность распределения соответствует уверенности алгоритма, а предсказаным положением объекта является математическое ожидание этой случайной величины. Алгоритм строит дискретную аппроксимацию распределения в виде набора взвешенных «частиц». Использование распределения положения вместо его моды позволяет найти сопровождаемый объект, даже когда на нескольких предыдущих кадрах объект локализован неверно.

Для качественного визуального сопровождения объекта важно построение его представления, описывающего особенности сопровождаемого объекта. Одним из используемых видов представления является набор локальных особенностей — ключевых точек — изображения объекта. При этом поиск на последующих кадрах происходит не всего объекта целиком, а только этих точек. В работе [24] было показано, что выбор углов текстуры объекта в качестве локальных особенностей позволяет добиться высокой устойчивости к потере объекта при сопровождении. Эти результаты подтверждаются работой [25], в которой предлагается использование нескольких таких точек и процедура их совместного отслеживания в видео.

Основным недостатком алгоритмов визуального сопровождения является необходимость качественной начальной инициализации положения объекта. В задаче сопровождения людей в видеопоследовательности для решения этой проблемы используют специализированные алгоритмы обнаружения, однако это приводит к появлению траекторий ложных срабатываний детектора. Также методы визуального сопровождения не учитывают наличие нескольких объектов сопровождаемого класса в видеопоследовательности. Это приводит к ошибкам двух видов: переключению сопровождения на другой схожий объект и построение нескольких траекторий для одного объекта.

1.3.2 Сопровождение через обнаружение

Подход сопровождения через обнаружение применяется для построения траекторий движения только заранее определенного класса объектов, например, машин или людей. При этом процесс обработки данных разбивается на этапы обнаружения объектов на ключевых кадрах видео и разбиения множества обнаружений на группы, соответствующие наблюдениям разных объектов.

Качество работы алгоритмов этого подхода зависит от используемого алгоритма обнаружения объектов на изображении и способа объединения обнаружений в траектории. Увеличение ложных срабатываний детектора повышает вероятность построения ложноположительных траекторий, не соответствующих объектам интереса. В работе [26] предполагают, что такие траектории соответствуют регулярным срабатываниям детектора на области фонового изображения, а поэтому являются неподвижными. Для борьбы с ложноположительными траекториями авторы используют статистику движения объекта на изображении. Если объект не был обнаружен ни на одном ключевом кадре, то в рамках рассматриваемого подхода для него не может быть построена траектория. Чтобы повысить полноту обнаружения людей, в том числе тех, чьё изображение частично перекрыто, в некоторых случаях детектор людей заменяют детекторами частей их тела [27].

Обычно задачу построения траекторий сводят к дискретной задаче разбиения множества обнаружений на группы (траектории). В работах [26—29] используется моделирование траектории движения динамической байесовской сетью и поиск максимума апостериорной вероятности. Свойство марковости таких сетей позволяет определять состояние объекта через его состояние в предыдущий момент времени. В работах [30; 31] авторы сводят задачу построения траекторий к задаче поиска потока минимальной стоимости в графе.

Альтернативный подход к сопровождению заключается в построении непрерывных кривых (траекторий), согласующихся с обнаруженными положениями объектов. В работах [32; 33] предлагается моделировать траекторию движения человека с помощью сплайнов. Авторы разбивают задачу на этапы ассоциирования обнаружений детектора с объектами и построения траекторий их движения. Построение траекторий заключается в циклическом повторении этих этапов в рамках минимизации общего функционала.

Большинство методов построения траекторий можно выразить в виде задачи минимизации функционала, ключевой фактор которого определяет насколько вероятно группе обнаруженийй соответствовать одному объекту. Этот

фактор описывает схожесть обнаружений внутри группы и основывается на решении задачи верификации. При сопровождении людей во многих случаях невозможно применить методы верификации человека по изображению лица, так как оно может быть скрыто или иметь низкое разрешение. Поэтому широко используется информация о положении людей в кадре и скорости их движения. Визуальное сопровождение применяется в алгоритмах сопровождения через обнаружение для оценки скорости движения объекта в окрестности кадра, где он был обнаружен [26; 27]. Эта информация позволяет сопоставить два результата обнаружения даже в случае, когда они соответствуют далёким (по времени) кадрам исходной последовательности. В работе [33] предлагают учитывать кривизну траектории и расстояние до других траекторий в видеопоследовательности в качестве регуляризации. Авторы [34] предсказывают направление движение людей в видео, используя семантическую информацию о сцене. В работах [29; 30] учитывается движение людей группами. Авторы оценивают траекторию всей группы и каждого человека внутри неё. В работах [35; 36] предлагается объединить обнаружение объектов на кадре и их сопровождение в видеопоследовательности в рамках единой задачи оптимизации.

Предложенный в этой работе метод относится к подходу сопровождения через обнаружение. В работе предлагается использовать информацию о калибровке камеры при обнаружении людей и учитывать области входа/выхода сцены при построении траекторий.

1.4 Определение позы человека

Существует несколько способов задания позы человека на изображении. В первых работах под позой понимали положение множества частей тела человека, таких как голень, предплечье, туловище, голова и др. Положение каждой части тела описывалось прямоугольником, стороны которого могли не

быть параллельны осям координат. Решение задачи в такой постановке оказалось сложным поскольку размеры частей тела на изображении могут сильно изменяться в зависимости от позы человека. Поэтому в работе [37] авторы определяют позу как положение суставов, соединяющих части тела человека, на изображении. В настоящее время эта постановка используется для определения позы.

1.4.1 Определение позы человека на изображении

Задача определения позы человека заключается в нахождении положения фиксированного множества точек на изображении человека. Некоторые из этих точек соответствуют суставам тела человека и образуют его виртуальный скелет. От задачи детектирования оценка позы отличается предсказанием структурированного выхода, то есть положения разных суставов на изображении зависит друг от друга. Эта зависимость определяется физическими размерами тела человека. Существует два основных метода определения позы человека на изображении: использование модели из набора деформируемых частей и регрессия положения суставов.

Модель из набора деформируемых частей

Первый из рассматриваемых подходов объединяет этапы локализации отдельных суставов скелета и выбора наиболее вероятной конфигурации позы человека в рамках общей задачи минимизации. Отличительная особенность данного метода заключается в возможности оценить правдоподобие любой позы

на рассматриваемом изображении. Другими словами, модель из набора деформируемых частей задает вероятностное распределение позы человека на изображении. Это свойство я активно использую в своей работе.

Модель из набора деформируемых частей является расширением стандартных методов локализации на случай объектов, состоящих из нескольких частей. Особенностью данного подхода является возможность учитывать допустимые изменения взаимного положения частей объекта относительно друг друга. Впервые этот подход был применен для обнаружения объектов в сцене [38], но в дальнейшем его применили для определении позы человека на изображении [37].

Объект моделируется марковской сетью, вершины которой соответствуют искомым суставам, а ребра задают ограничения на их взаимное расположение. В работах [37; 39] авторы ограничиваются рассмотрением только парных потенциалов, заданных на ребрах. Чтобы вывод в графической модели был точным и эффективным, авторы ограничиваются рассмотрением только моделей в виде дерева.

В общем виде модель из набора частей определяет позу P человека как минимум функции энергии, описываемой двумя типами потенциалов (факторов):

$$E(P) = \sum_{i \in V} \varphi_i(p_i, s) + \sum_{(i,j) \in E} \psi^s_{(i,j)}(p_i, p_j, s), \tag{1.1}$$

где φ_i — унарный потенциал сустава p_i, $\psi^s_{(i,j)}$ задаёт парный потенциал для суставов p_i и p_j на изображении, а s — дискретный параметр размера человека. В работах [37; 39] предполагается, что парный потенциал ψ_i не зависит от входного изображения. Благодаря использованию глобального скрытого параметра размера s тела человека, удается избежать ситуаций, когда в найденной позы одни части непропорционально больше других.

Одним из недостатков модели из набора деформируемых частей, предложенной в [37], является древовидная структура зависимостей между суставами. Например, положения коленей не имеют прямой зависимости, и связаны через

положение суставов туловища. Это приводит к тому, что алгоритм может расположить суставы обеих ног человека на изображении одной ноги. Для решения этой проблемы в работе [40] предлагается расширить модель человека набором *позлетов* (в английской версии *poselet*), которые ограничивают взаимное расположение некоторого подмножества суставов тела человека. Несмотря на то, что полученная графическая модель больше не является деревом, алгоритм вывода остаетсяся эффективным, поскольку допускает перебор по небольшому множеству состояний позлетов.

Базовая модель [37] предполагает, что все суставы тела человека видны на изображении. Это становится серьезной проблемой в ситуациях частичной видимости тела человека на изображении из-за перекрытий и частичным выходом человека за границу изображения. В работах [41; 42] авторы определяют является ли сустав перекрытым изображением другого человека.

В работе [43] указывается важность моделирования внешности человека для более точной оценки его позы. Авторы расширили модель позы человека гистограммой цветов его изображения и предложили метод совместного оценки параметров, который повышает точность решения задачи определения позы.

В работах [37; 39—41] авторы используют эвристические признаки для описания изображения. Обучение параметров графической модели происходит с помощью структурного метода опорных векторов [44]. В работе [45] авторы показывают, что вывод в модели из набора деформируемых частей может быть представлен в виде сверточной нейронной сети прямого распространения. Это позволило в последующих работах [42; 46; 47] обучать признаки изображения совместно с параметрами графической модели.

В работе [48] авторы показали, что модель из набора деформируемых частей позволяет находить не только одну позу, минимизирующую $E(P)$, но также искать и другие минимумы функционала, отличающиеся в положении хотя бы одного сустава от предыдущих. Множество таких минимумов описывает множество гипотез наилучшых поз человека на изображении. В случае, когда количество необходимых гипотез значительно меньше количества возможных

положений одного сустава на изображении, вычислительная сложность построения множества таких гипотез не более чем в два раза превосходит сложность построения наилучшей гипотезы позы человека.

В диссертационной работе я использовал алгоритм определения позы на изображении, основанный на модели из набора деформируемых частей. Возможности построения гипотез позы человека на изображении и расширения минимизируемого функционала позволили мне применить его для определения позы человека в видеопоследовательности.

Регрессия положения суставов

Альтернативным подходом к определению позы человека на изображении является метод регрессии положения суставов из изображения. В отличие от модели из набора деформируемых частей он не позволяет оценить качество произвольной позы на изображении, но может неявно учитывать положения групп суставов.

В работе [49] авторы построили отображение входного изображения в координаты каждого сустава. Предложенный ими алгоритм представляет каскад из двух нейронных сетей, которые последовательно уточняют положение каждого сустава на изображении. Авторы показали, что первый регрессор указывает приближенное положение суставов всего тела, в то время как второй, уточняет положение отдельных суставов. В работе [50] авторы указали, что предсказание тепловой карты положения суставов позволяет добиться лучших результатов в рамках того же подхода.

В работе [51] авторы расширили предыдущих подход. Для уточнения локализации каждого сустава они предложили использовать также тепловые карты положения других суставов. Такой подход наиболее близок к модели из набора

деформируемых частей, но позволяет неявно учитывать положение всех суставов на изображении при их уточнении. Наиболее сложной для данного метода является ситуация наличия на одном изображении нескольких людей, изображения которых частично перекрывают друг друга.

1.4.2 Определение позы человека в видеопоследовательности

Модель из набора деформируемых частей допускает расширение на случай последовательности кадров. Для этого вводится модель движения, описывающая изменение позы между кадрами. Наиболее простым её вариантом является задание независимых моделей движения для каждого сустава:

$$E(\{P_t\}_{t=1}^{T}) = \sum_{t=1}^{T} E(P) + \sum_{i \in V} \sum_{t=1}^{T-1} \psi_i^s(p_i^{t+1}, p_i^t, s^t), \tag{1.2}$$

где $E(P)$ — модель позы человека на изображении (1.1).

В работе [48] предлагается простая модель изменения позы, предполагающая слабое её изменение между кадрами. В качестве парного потенциала выбирается квадратичная функция изменения положения суставов между кадрами.

Поиск минимума функционала (1.2) оказывается сложной задачей, так как соответствующая графическая модель содержит циклы. Точный вывод оказывается невозможным из-за его высокой вычислительной сложности. Поэтому авторы [48] использовали построение наилучших гипотез позы человека на изображении, чтобы уменьшить количество допустимых поз и свести задачу определению оптимального состояния в марковской цепи.

В диссертационной работе я предлагаю обобщение минимизируемого функционала, учитывающее скорости движения каждого сустава скелета человека. Также предлагается алгоритм поиска его локального минимума, как по множеству положений суставов, так и по параметрам их скорости.

Глава 2. Определение позы камеры

В данной главе я предлагаю метод определения позы камеры, основанный на анализе размеров изображений объектов интереса в наблюдаемой сцене. Предложенный метод имеет два ключевых преимущества:

1. он допускает наличие ложных обнаружений объектов интереса и учитывает погрешность локализации присутствующих объектов;
2. он предсказывает положение и направление камеры камеры на диапазоне угла её наклона $\left(0, \frac{\pi}{2}\right)$.

Позой камеры называется её положение и направление относительно снимаемой сцены. Таким образом, задача определения позы камеры заключается в поиске преобразования мировой системы координат в систему координат, связанную с камерой. Для формализации постановки задачи в следующем разделе я представляю математическую модель наблюдаемых данных и используемой камеры.

2.1 Математическая модель наблюдаемых данных

В данной работе я предлагаю модель наблюдаемых данных, состоящую из плоской статичной сцены и людей, находящихся в ней. Такая аппроксимация подходит для описания большинства сценариев видеонаблюдения. Ниже предлагается описание трёх её составляющих: сцены, камеры и человека.

2.1.1 Модель сцены

Под сценой в данной работе я подразумеваю неподвижные объекты, изображения которых получает камера. Таким образом сцена может состоять из дорог, зданий, деревьев, скамеек и др. Однако предлагаемый метод не использует семантическую информацию об объектах сцены, поэтому в данной работе я рассматриваю простейшую модель сцены, состоящую из единственной горизонтальной плоскости — плоскости земли.

Для задания позы камеры необходимо выбрать мировую систему координат, связанную со сценой. Пусть x_w, y_w и z_w обозначают её базисные векторы. Мировая система координат выбрана таким образом, что плоскость земли совпадает с плоскостью $z = 0$, а вектор z_w совпадает с направлением вверх в сцене. В качестве начала мировой системы координат я выбрал проекцию положения камеры на плоскость земли. В данной работе я предполагаю, что вектор x_w коллинеарен проекции вектора направления камеры на плоскость земли. Описанные ограничения однозначно определяют мировую систему координат в наблюдаемой сцене.

2.1.2 Модель камеры

С камерой также связанна система координат. Для обозначения её базисных векторов в данной работе я использую x_c, y_c и z_c. Оптический центр камеры задает начало её координат, x_c совпадает с направлением вправо на изображении, а y_c — с направлением вниз. Вектор z_c указывает направление камеры.

Я использую модель перспективной проекции камеры, которая описывается фокусным расстоянием камеры f_c. Физические характеристики используемой камеры задаются несколькими параметрами: размером матрицы камеры

w_c, h_c, положением принципиальной точки (x_c, y_c), размерами пикселя (w_p, h_p) и углом его скоса α_c. Я использую предположения о квадратной форме пикселей ($w_p = y_p$, $\alpha_c = 0$) и положении принципиальной точки в центре изображения ($x_c = \frac{w_I}{2}$, $y_c = \frac{h_I}{2}$).

При заданных ограничениях модель перспективная проекция полностью определяется матрицей внутренней калибровки камеры, которая имеет следующий вид:

$$K = \begin{bmatrix} f & 0 & \frac{w_I}{2} \\ 0 & f & \frac{h_I}{2} \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.1}$$

где через $f = \frac{f_c}{w_p}$ — фокусное расстояние, вычисленное в размерах пикселя, а (w_I, h_I) — размер изображения.

Несмотря на простоту, используемая модель камеры достаточна для описания и более общих случаев. Например, если для некоторой камеры известны параметры дисторсии и положения приципиальной точки, то её можно привести в соответствие с описанной моделью путем применения детерминированной процедуры модификации каждого кадра входной видеопоследовательности.

2.1.3 Модель человека

Единственными движущимися объектами в сцене являются люди. Я использую модель человека, предложенную в статье [52]. Она задаёт отображение параметров позы и комплекции в положение вершин трехмерной модели человека.

2.2 Поза камеры

Выбранная модель сцены однозначно определяет взаимное положение мировой системы координат и системы координат камеры. При заданных ограничениях поза камеры l_c однозначно определяется тремя параметрами:

- высотой h камеры над плоскостью земли;
- углом t наклона камеры;
- углом r крена камеры.

Высота h камеры над плоскостью земли определяет положение оптического цента камеры на оси z. Углы наклона t и крена r камеры являются углами нутации и собственного вращения при преобразовании мировой системы координат в систему координат камеры.

Формально задача определения позы камеры по изображению имеет следующий вид:

Вход:
- Последовательность $\{I_t\}_t^T$ изображений, полученная статичной камерой;
- Фокусное расстояние f камеры;

Выход: Параметры позы камеры в сцене: $l_c = (h, t, r)$.

Разработанный алгоритм может применяться для последовательностей, содержащих как цветные, так и монохромные изображения. На наблюдаемые данные накладываются следующие ограничения:

- в данный представлено не менее трех изображений людей, не расположенных на одной прямой;
- изображения голов людей имеют размер не менее 16×16 пикселей;
- высота h камеры над землей не превосходит 20 метров;
- угол крена r камеры находится в пределах $\left(-\frac{\pi}{6}, \frac{\pi}{6}\right)$;
- угол наклона камеры t находится в пределах $\left(0, \frac{\pi}{2}\right)$;
- фокусное расстояние f камеры ограничено 5000 размерами пикселя.

Таблица 1 — Распределение параметров позы и фокусного расстояния камеры в синтетической выборке.

Параметр	Название	Минимальное значение	Максимальное значение
h	высота (м)	0	20
t	наклон (рад)	0	$\frac{\pi}{2}$
r	крен (рад)	$-\frac{\pi}{6}$	$\frac{\pi}{6}$
f	фокусное расстояние (пиксели)	0	5000

Входная последовательность изображений может иметь произвольную длину, в частности содержать единственное изображение.

2.3 Предложенный метод

Я разработал метод определения позы камеры, основанный на анализе размеров изображений объектов интереса. В качества объектов были выбраны головы людей, поскольку в отличие от фигур всего человека они меньше подвержены перекрытиям в сценарии видеонаблюдения.

Для построения отображения входного изображения в параметры камеры я использовал методы машинного обучения с учителем. В связи с отсутствием больших размеченных коллекций с известными положениями объектов и камеры в сцене обучение происходило на синтетической выборке.

2.3.1 Построение обучающей выборки

Обучающая выборка состоит из синтетических последовательностей изображений. Для её построения я использовал модель наблюдаемых данных,

описанную в разделе 2.1. Параметры позы камеры выбирались равномерно из диапазона, указанного в таблице 1.

Предложенный алгоритм оценки позы камеры использует только результаты обнаружения людей, то есть положение и размер головы человека на изображении. Поэтому в синтетической выборке не моделируется разнообразие поз людей реального мира, и все люди находятся в стандартной позе. В качестве роста людей в сцене я выбрал 1.75 метра — средний рост взрослых европейцев.

При построении синтетической выборки отбраковывались изображения, не удовлетворяющие ограничениям модели наблюдаемых данных. Построенная синтетическая выборка состоит из 100373 последовательностей, каждая из которых содержит не менее 200 изображений. На одном изображении запечатлен лишь один человек, расположенный в произвольном месте изображения. Отсутствие перекрытий позволяет обнаружить человека детектором и выявить его ложные срабатывания.

2.3.2 Выбор признакового описания

Синтетические последовательности изображений визуально отличаются от реальных данных (рис. 2.1). Поэтому для обучения алгоритма регрессии позы необходимо выбрать признаковое описание инвариантное к используемой выборке.

В качестве описания я использую результаты обнаружения голов людей на изображении. Таким образом, каждый человек на изображении описывается тройкой чисел (x_h, y_h, s_h), соответствующих положению центра его головы и её линейному размеру. Конечно, модель человека [52] позволяет определить истинное положение головы человека на синтетическом изображении. Однако, такой способ не может быть применен к изображения реальных данных. Использовании разных способов оценки положения головы человека на реальных

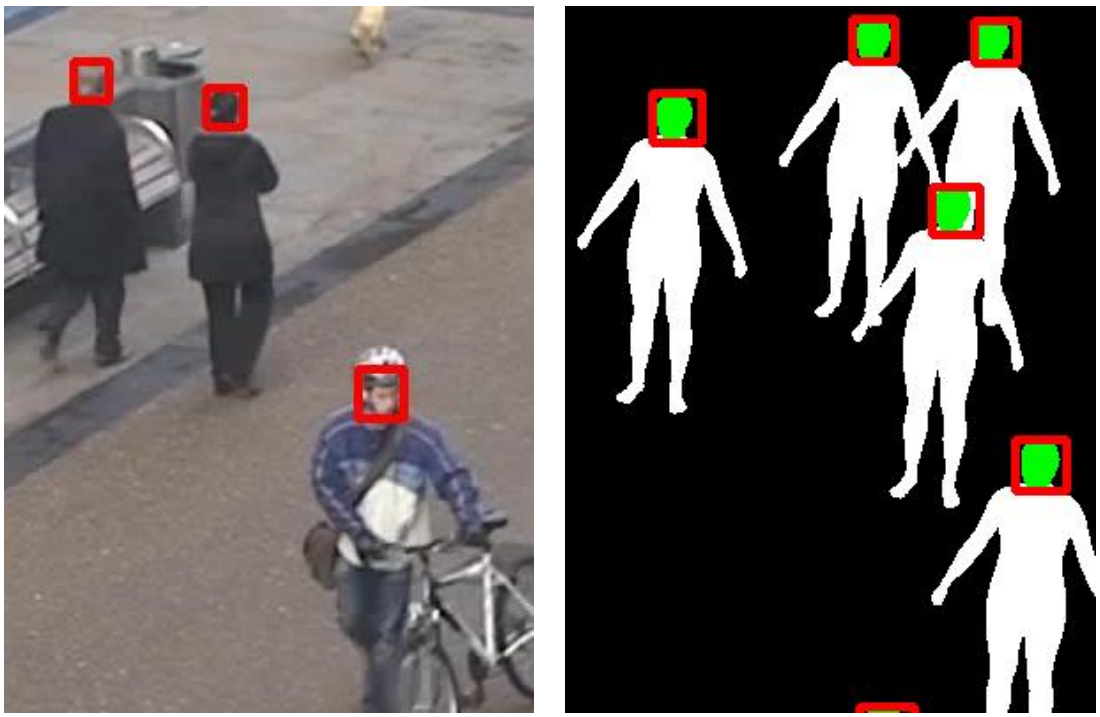


Рисунок 2.1 — Сравнение наблюдаемого кадра видеопоследовательности и синтетического. Красным выделены области головы экспертной разметки (для реального кадра) и обнаруженные детектором (для синтетических данных).

и синтетических данных может привести к уменьшению обобщающей способности алгоритма определения позы камеры. Поэтому детектор головы человека применяется даже к синтетическим данным. Более того, такой подход позволяет в процессе обучения адаптироваться к погрешности в локализации и определении размеров головы человека, совершаемые алгоритмом детектирования. Также этот подход позволяет использовать для определения позы камеры любые объекты, для которых доступна трехмерная модель и алгоритм обнаружения на изображении.

Для локализации голов людей на изображении был выбран алгоритм [53] и использована его оптимизированная версия[1]. Использование данного алгоритма обусловлено высокой скоростью его работы. Этот фактор оказался очень важным для построения большой коллекции изображений синтетической выборки. Также данный алгоритм не чувствителен к отсутствию текстуры на изображении человека.

[1]https://bitbucket.org/13e_sha/fasterhog

Для построения прецедента для обучения алгоритма регрессии позы камеры я использовал информацию о 64 людях в сцене.

2.3.3 Регрессия позы камеры

Рассматриваемая задача отображения размеров найденных людей в параметры позы камеры является задачей регрессии.

При обучении регрессии ключевым является выбор оптимизируемой функции потерь. Стандартным выбором функцией ошибки является Евклидово расстояние между предсказанными и истинными значениями целевой переменной. Оно одинаково «штрафует» отклонение от правильного ответа для всех прецедентов. Однако, при увеличении высоты камеры над плоскостью земли, точность её предсказания может уменьшаться. Например, изменение высоты камеры на одно и то же значение может не привести к заметным изменениям наблюдаемых данных, если камера установлена высоко, либо существенно изменить распределение размеров голов людей на изображении, если камера расположена вблизи плоскости земли. Поэтому точность предсказания должна зависеть от позы камеры. Другими словами, разные последовательности выборки могут иметь различную сложность при обучении.

Я учитываю это предположение в минимизируемой функции потерь. Она представляет взвешенную сумму квадратичных отклонений, где и значение целевой переменной, и вес его вклада в функцию потерь предсказывается алгоритмом регрессии.

Формально, построенный алгоритм использует предположение нормального распределения позы камеры $l_c = (h, t, r)$ при условии наблюдаемого прецедента x и предсказывает математическое ожидание $\tilde{l_c}$ и дисперсию Σ_c этого распределения:

$$p(l_c|x, \Theta) = N(l_c|\tilde{l_c}(x, \Theta), \Sigma_c(x, \Theta)), \tag{2.2}$$

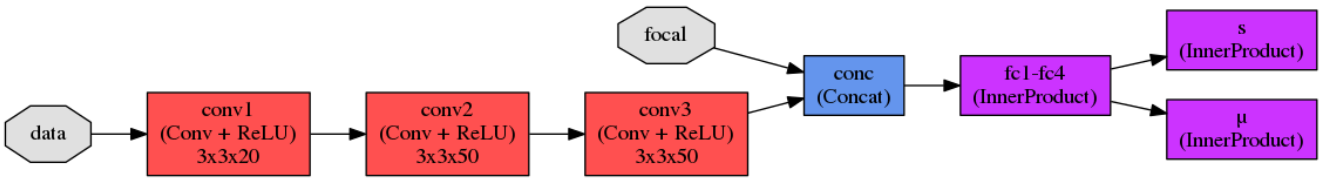


Рисунок 2.2 — Схема нейронной сети для предсказания параметров положения и направления камеры.

где Θ — обучаемые параметры.

Матрица ковариации Σ_c должна быть положительно определенной. Я учитываю это ограничение в функции потерь, рассматривая только диагональные матрицы ковариации следующего вида:

$$\Sigma_c(x, \Theta) = diag\left(e^{s(x,\Theta)}\right) + \varepsilon I, \tag{2.3}$$

где $s(x, \Theta) = (s_h(x, \Theta), s_t(x, \Theta), s_r(x, \Theta))$ — вектор параметров матрицы ковариации, предсказанный нейросетью. Чтобы матрица ковариации не становилась вырожденной в процессе обучения, к ней добавляется параметр регуляризации ε, равный 10^{-6}.

Определитель матрицы ковариации Σ_c описывает размер области неопределенности предсказания позы камеры. Поэтому величину $\lambda_c(x, \Theta) = \left|\Sigma_c^{-1}(x, \Theta)\right|$ можно интерпретировать как уверенность алгоритма в предсказании.

Обучение отображений $\tilde{l_c}(x, \Theta)$ и $\Sigma_c(x, \Theta)$ производится с помощью метода максимального правдоподобия. Таким образом, в качестве функция потерь используется отрицательный логарифм правдоподобия наблюдаемых данных:

$$L\left(\left\{l_c^i\right\}_i | \left\{\tilde{l_c^i}\right\}_i, \left\{s^i\right\}_i\right) = -\sum_i \log N\left(l_c^i | \tilde{l_c^i}, diag\left(e^{s^i}\right) + \varepsilon I\right) \tag{2.4}$$

Производные используемой функции потерь могут быть вычислены аналитически:

$$\frac{\partial L}{\partial \tilde{l_c^j}} = \frac{\tilde{l_c^j} - l_c^j}{e^{s^j} + \varepsilon} \tag{2.5}$$

$$\frac{\partial L}{\partial s^j} = \frac{1}{2}\frac{e^{s^j}}{e^{s^j} + \varepsilon}\left(1 - \frac{(\tilde{l_c^j} - l_c^j)^2}{e^{s^j} + \varepsilon}\right) \tag{2.6}$$

Выражения (2.4), (2.5) и (2.6) допускают эффективную реализацию используемой функции потерь на современных графических ускорителях.

Я использую сверточную нейронную сеть прямого распространения, показанную на рисунке 2.2, в качестве функций $\tilde{l_c}(x, \Theta)$ и $\Sigma_c(x, \Theta)$. Входом нейронной сети является тензор размера $3 \times 8 \times 8$, описывающий положение и размер каждой из 64 обнаруженных голов. Чтобы алгоритму не потребовалось адаптироваться ко всевозможным перестановкам объектов прецедента, они были отсортированы по возрастанию размера. Дополнительным входом сети является фокусное расстояние f используемой камеры, вычисленное в размерах пикселя изображения.

Построенная нейронная сеть состоит из 3 сверточных слоёв, после которых используется нелинейная функция ReLu. Размер каждой сверки равер 3×3. Такой подход позволяет сверточным слоям 1) использовать информацию об удаленных объектах, размер которых существенно отличается и 2) адаптироваться к шуму в данных за счет использования объектов, чьи размеры близки.

После применения сверточных слоёв полученный результат объединяется с фокусным расстоянием f камеры и подается на вход полносвязным слоям. Описанная выше функция потерь позволяет обучать нейронную сеть предсказывать положение камеры и уверенность в предсказании.

2.3.4 Объединение результатов прецедентов

Построенная нейронная сеть предсказывает позу камеры, используя положения не более 64 людей в сцене. В реальных сценариях видеонаблюдения количество людей в сцене может существенно превышать это значение. Поэтому ниже я рассматриваю метод объединения результатов на разных наборах данных.

Объединить результаты работы алгоритма на K разных подмножествах обнаружений людей можно с помощью наивного Байессовского метода:

$$\overline{l_c} = \overline{\Sigma_c} \left(\sum_{k=1}^{K} \Sigma_{c,k}^{-1} \tilde{l}_{c,k} \right) \tag{2.7}$$

$$\overline{\Sigma_c} = \left(\sum_{k=1}^{K} \Sigma_{c,k}^{-1} \right)^{-1}, \tag{2.8}$$

где $\overline{l_c}$ — предсказанная поза камеры.

Важно отметить, что данных подход предполагает отсутствие зависимости между предсказанными позами камеры на разных подмножествах данных. Чтобы этого добиться, можно использовать непересекающиеся подмножества обнаружений объектов на разреженном подмножестве кадров.

2.4 Обучение и экспериментальная оценка

В этом разделе описывается используемый метод обучения параметров нейронной сети и результаты тестирования на синтетических и реальных данных.

2.4.1 Обучение

Обучение нейронной сети проводится на построенной синтетической выборке, содержащей только корректные обнаружения голов людей. Использование только таких «чистых» данных не позволяет построенному алгоритму обобщаться к результатам работы детектора на реальных данных, содержащие ошибки.

Для решения этой проблемы я предлагаю два метода моделирования шума в реальных данных: 1) ложные обнаружения голов в сцене; 2) дублирующиеся

обнаружения в прецеденте. Второй тип шума может возникать в реальных данных, если наблюдаемых человек не меняет своего положения в сцене в течение некоторого времени. Для моделирования зашумленных данных применяется следующий алгоритм построения прецедентов:

1. из множества обнаружений, относщихся к одной последовательности, выбирается подмножество из n элементов $(0 < n \leqslant 64)$;
2. среди выбранных обнаружений m $\left(0 < m < \frac{n}{10}\right)$ произвольных заменяются на ложные срабатывания детектора со случайным положением и размером на изображении;
3. из построенного множества генерируется прецедент с помощью выборки 64 обнаружений с повторениями.

Это позволяет моделировать ложные срабатывания алгоритма детектирования людей и дублирование обнаружений на разных кадрах. С помощью предложенного алгоритма для каждой последовательности синтетической выборки происходит построение трех прецедентов.

Предложенная сверточная нейронная сеть имеет всего 67393 параметра и относительно небольшое количество нейронов в скрытых слоях. Это позволяет при обучении осуществлять пакетную обработку данных с пакетом, содержащим 32768 прецедентов. Для обучения я применяю метод Adam [54]. Скорость обучения понижается по степенному закону с параметром $\boldsymbol{\gamma} = 0.95$ каждые 1500 итераций. Обучение занимает 200000 итераций. В экспериментах я использую 80% последовательностей синтетической выборки для обучения и 20% для валидации.

2.4.2 Экспериментальная оценка на синтетической выборке

Экспериментальная оценка содержит результаты применения предложенного алгоритма к синтетическим и реальным данным видеонаблюдения.

Таблица 2 — Зависимость нормализованного отклонения на обучающей и валидационной выборках от количества полносвязных слоёв и размера обучающей выборки.

Размер выборки	Количество полносвязных слоёв	Средняя ошибка на обучении	Средняя ошибка на валидации
20448	3	0.1061	0.121
20448	4	0.0956	0.1167
20448	5	0.07	0.12
30179	4	**0.09015**	0.117
51366	4	0.09836	0.1064
80298	5	0.09764	**0.1009**

Первый эксперимент показывает влияние размера обучающей выборки и количество скрытых слоев нейронной сети на качество построенного регрессора. Я провел тестирование нейронных сетей с разным количеством полносвязных слоёв (см. таблицу 2). Среднеквадратичная ошибка регрессора на валидационной выборке оказывается неинформативной мерой качества алгоритма, так как размерности и диапазоны значений параметров позы камеры различаются. Поэтому я использую L_1 расстояние между предсказанными и истинными значениями параметров позы камеры, нормализованными на диапазон их значений в обучающей выборке (см. таблицу 1). Результаты сравнения показывают, что увеличение размера обучающей выборки и количество полносвязных слоёв приводит к повышению точности определения позы камеры. Наилучших результатов удаётся добиться при обучении нейронной сети, содержащей 5 полносвязных слоёв, на обучающей выборке из 80298 сцен. Также тестирование показывает, что при выбранных параметрах не происходит переобучение, так как ошибки на обучающей и валидационной выборках близки.

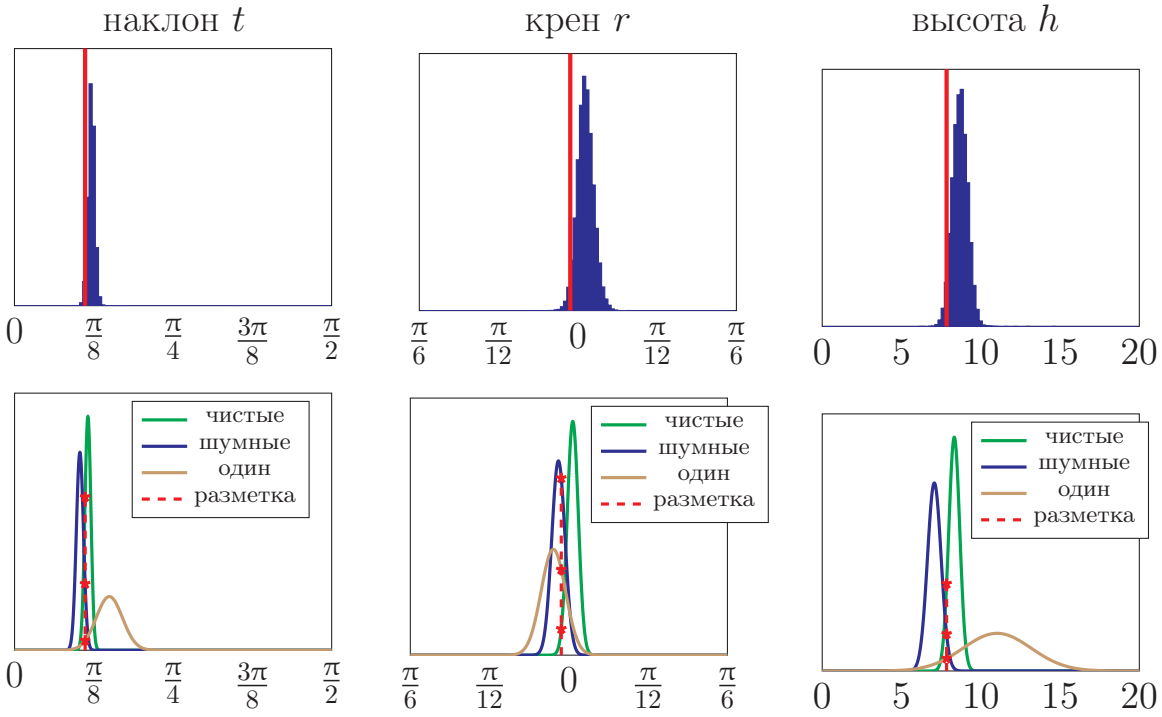


Рисунок 2.3 — Результаты определения позы камеры на выборке TownCentre. В первой строке представлены гистограммы предсказанных параметров позы камеры на разных подмножествах корректных обнаружений людей в выборке (голубым). Вторая строка описывает предсказанные распределения параметров позы камеры, полученные при объединении результатов предсказаний на а) «чистых» данных, не содержащих шума (зеленая кривая); б) «шумных» данных, содержащих ложные срабатывания детектора человека (синяя кривая); в) прецедентах, содержащих 64 копии обнаружения одного человека (коричневая кривая). Параметры позы камеры, представленные в экспертной разметке, отмечены красным. Столбцы соответствуют углу наклона, углу поворота и высоте камеры над плоскостью земли.

2.4.3 Экспериментальная оценка на реальных данных

Для оценки обобщающей способности алгоритма необходимо также представить результаты тестирования на доступных выборках реальных данных видеонаблюдения. Эксперименты включают в себя оценку качества на выборке TownCentre [26], так как она содержит экспертную разметку положения голов людей и калибровки камеры.

Таблица 3 — Результаты определения позы камеры на выборке TownCentre

	Наклон t	Поворот r	Высота h
экспертная разметка	0.3497	-0.0251	7.84
«чистые» данные	0.3634 ± 0.0149	0.0130 ± 0.0182	8.3290 ± 0.3453
«зашумлённые» данные	0.3237 ± 0.0176	-0.0345 ± 0.0219	7.0696 ± 0.4294
единственное обнаружение	0.4694 ± 0.0649	-0.0513 ± 0.0402	11.0312 ± 2.1441

При тестировании к кадрам последовательности применяется алгоритм детектирования голов людей на изображении fastHOG [53]. Корректными считаются те обнаружения, чьё пересечение с результатами экспертной разметки превосходит значение 0.5 по метрике IoU. В таких условиях точность алгоритма составляет 48%, и результаты детектирования содержат 19061 обнаружение головы человека на 4501 кадре. Для тестирования из них случайным образом выбирается 40000 прецедентов, содержащих по 64 обнаружения с повторениями. В первой строке рисунка 2.3 показаны распределения ответов на разных прецедентах. Можно заметить, что указанные распределения имеют моду близкую к правильному значению параметров.

Для объединения результатов работы алгоритма на разных прецедентах используется подход, описанный в разделе 2.3.4. Случайным образом выбирается 20 прецедентов, множества обнаружений которых не пересекаются. Это обеспечивает независимость результатов работы алгоритма на разных прецедентах. При объединении предсказаний оцениваются оба параметра: среднее здачение целевой переменной и её матрица ковариации. Более того полученная матрица ковариации также имеет диагональный вид. Поэтому результаты объединения предсказаний на разных прецедентах можно представить в виде функции плотности распределения отдельных компонет позы камеры (см. рисунок 2.3 вторая строка зеленая кривая).

На рисунке 2.4 представлена визуализация синтезированных людей на плоскости земли при предсказанной позе камеры. Можно заметить, что

Рисунок 2.4 — Визуализация синтезированных людей на предсказанной плоскости земли.

размеры реальных и синтезированных людей похожи, что подтверждает правдоподобие предсказанной позы камеры.

В следующем эксперименте предложенный метод определения позы камеры повторяется, используя все обнаружения на выборке TownCentre, в том числе и ложноположительные срабатывания детектора. Полученные результаты представлены во второй строке строке рисунка 2.3 (синяя кривая). Важно отметить, что доля ложных срабатываний детектора в выборке существенно превосходит моделируемую долю ложных обнаружений в обучающей выборке (52% против 10%). Тем не менее результаты показали, что предсказанная поза камеры близка к экспертной разметке даже в ситуации наличия большого количества ложных срабатываний детектора.

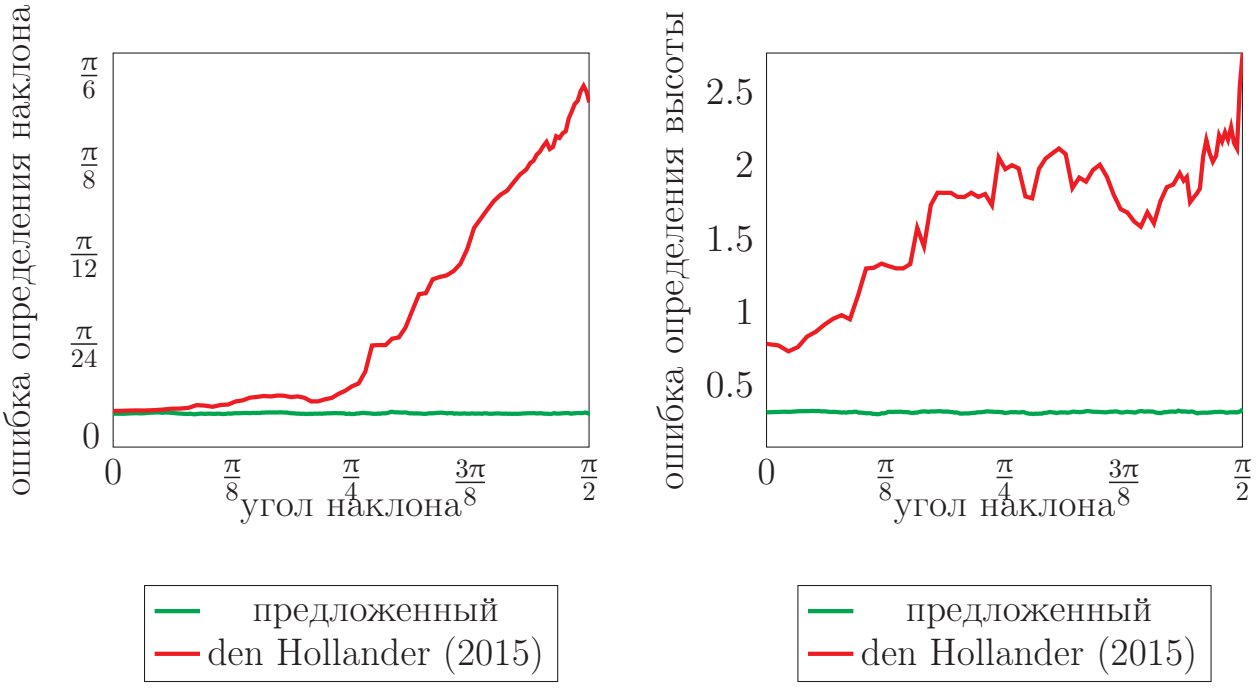
В последнем эксперименте с выборкой TownCentre рассматривается экстремальный случай наличия повторяющихся обнаружений в прецеденте. В данном случае прецедент состоит из 64 копий одного случайно выбранного обнаружения головы человека в сцене. Такая ситуация может возникать в случае, когда в сцене присутствует лишь один человек, который не меняет своего положения в течении длительного времени. Алгоритм не может предсказать положение камеры, так как единственное обнаружения задает лишь расстояние от камеры до плоскости земли в одной точке. Во второй строке рисунка 2.3

Таблица 4 — Предсказанные параметры позы камеры на выборке PETS 2006. Таблица содержит предсказанные параметры позы камеры и их среднеквадратичные отклонения. В последней строке указана средняя ошибка на синтетической валидационной выборке.

Последовательность		Наклон	Поворот	Высота
PETS 1	разметка	0.104	-0.017	1.878
	оцененное	0.445 ± 0.061	-0.001 ± 0.035	4.595 ± 0.878
PETS 2	разметка	-0.037	-0.095	4.609
	оцененное	0.16 ± 0.071	-0.037 ± 0.042	9.37 ± 1.081
PETS 3	разметка	0.289	-0.03	5.501
	оцененное	0.357 ± 0.024	-0.024 ± 0.022	5.823 ± 0.298
PETS 4	разметка	0.458	-0.109	6.567
	оцененное	0.453 ± 0.024	0.059 ± 0.023	5.367 ± 0.326
Синтетическая	средняя ошибка	0.084	0.093	0.818

(коричневая кривая) представлена оценка поза камеры. Можно заметить, что дисперсия распределений существенно увеличивается при переходе к такому сложному прецеденту. Таким образом, предсказанный параметр точности алгоритма λ_c позволяет определить сложные для анализа прецеденты. Результаты всех экспериментов на выборке TownCentre представлены в таблице 3.

Также предложенный метод был протестирован на четырех последовательностях более сложной выборки PETS 2006 [55]. Важно отметить, что первая и вторая последовательность этой выборки нарушают одно из используемых предположений. На этих последовательностях люди расположены на нескольких этажах, следовательно не удается выделить единственную плоскость земли. Тем не менее, я применил предложенный метод ко всем последовательностям выбоки и использовал все обнаружения голов для построения прецедентов. Результаты тестирования (см. таблицу 4) показали, что предложенный метод верно оценивает позу камеры для третьей и четвертой последовательностей и



(а) (б)

Рисунок 2.5 — Зависимость ошибки предсказания наклона (а) и высоты (б) камеры от истинного значения угла наклона. Красным показана средняя ошибка алгоритма [7], а зеленым – средняя ошибка прдложенного алгоритма. Тестирование проводилось на синтетической тестовой выборке, содержащей только корректные срабатывания детектора детектора.

ошибается для первых двух последовательностей. Однако, для этих сложных последовательностей уверенность алгоритма оказывается существенно меньше.

В ходе анализа алгоритма также исследовалась зависимость точности определения позы камеры от угла её наклона. Для этого интервал значений угла наклона камеры $\left(0, \frac{\pi}{2}\right)$ был разбит на сегменты так, чтобы каждому сегменту соответствовало 500 примеров тестовой выборки. Внутри каждого сегмента была вычислена средняя ошибка определения угла наклона и высоты камеры двумя алгоритмами: предложенным и методом [7] (рис. 2.5). Поскольку в работе [7] предполагается отсутствие ложных срабатываний детектора, то при сравнении синтетическая выборка содержала только корректные обнаружения людей. Анализ показал, что точность определения положения камеры предложенным алгоритмом не зависит от значения угла наклона камеры. Причина такого поведения функции ошибки заключается в используемом при обучении

наборе данных. В процессе обучения нейронной сети приходится адаптироваться к всевозможным положениям и направлениям камеры в сцене. В то же время средняя ошибка оценки положения камеры в сцене методом [7] увеличивается с увеличеним угла наклона камеры. Причиной такого поведения является предположение, что границы ограничивающего прямоугольника содержат верхнюю и нижнюю точки детектируемого объекта. Это предположение нарушается тем больше, чем больше угол наклона камеры приближается к $\frac{\pi}{2}$.

2.5 Заключение

В этой главе предложен алгоритм определения положения и направления неподвижной камеры видеонаблюдения, использующий изображения людей в качестве калибровочных объектов. Отличительной особенностью построенного алгоритма является применение машинного обучения для построения отображения результатов детектирования людей в значения параметров позы камеры. Благодаря такому подходу, предложенный алгоритм обладает двумя важными свойствами:

1. устойчивостью к изменению угла наклона камеры в интервале $\left(0, \frac{\pi}{2}\right)$
2. возможностью предсказывать ошибку оценки положения камеры.

Результаты главы были опубликованы в [56].

Глава 3. Локализация людей на изображении

В данной главе рассматривается задача обнаружения людей на изображении, полученном камерой с известными параметрами калибровки. Внутренние свойства камеры, такие как фокусное расстояние, и её поза накладывают ограничения на возможные размеры и положения объектов на изображении. Например, в системах видеонаблюдения изображения людей не могут полностью находиться выше уровня горизонта. Использование таких ограничений позволяет повысить точность и скорость обработки данных по сравнению с базовыми методами детектирования.

В этой главе я предлагаю способ построения алгоритма детектирования людей на изображении, полученном камерой с известными параметрами калибровки. Предложенный метод учитывает ограничения, накладываемые положением камеры в сцене, для повышения точности и производительности базового алгоритма детектирования изображения головы человека. Также он допускает обобщение на случай обнаружения других классов объектов интереса в кадре, таких как автомобиль, пешеход и др.

Формально, рассматриваемая задача определяется следующим образом:

Вход:
1. Изображение I;
2. Параметры позы l_c камеры (см. главу 2);
3. Фокусное расстояние f камеры.

Выход: Множество прямоугольников, ограничивающих изображения объектов интереса.

3.1 Предложенный метод

Предложенный алгоритм представляет суперпозицию $g(A_b, A_f)$ базового алгоритма обнаружения изображения головы человека A_b и классификатора его результатов A_f. Предложенная суперпозиция не накладывает ограничений на выбор алгоритма A_b и использует его как «чёрный ящик». Результатом предложенного отображения $g(A_b,A_f)$ является множество только тех обнаружений алгоритма A_b, которые были классифицированы A_f как «правдоподобные» при заданных параметрах калибровки камеры.

У предложенной суперпозиции есть несколько важных свойств. Во-первых, за счет выявления «невозможных» для заданной сцены срабатываний детектора удается повысить точность обнаружения по сравнению с базовым алгоритмом A_b. Более того, если алгоритм A_b основан на методе скользящего окна, то классификатор A_f позволяет определить те окна, которые не содержат «правдоподобные» обнаружения, ещё до анализа изображения. Это свойство позволяет повысить скорость работы суперпозии $g(A_b, A_f)$ в системах видеонаблюдения, с неподвижной камерой.

Задача разработки алгоритма обнаружения $g(A_b, A_f)$ сводится к построению классификатора A_f. Для этого я использую метод машинного обучения, применённый к синтетическим данным видеонаблюдения. Таким образом, необходимо решить 3 задачи:

1. Построить синтетическую выборку наблюдаемых данных;
2. Построить признаки, инвариантные для синтетических и реальных данных;
3. Обученить классификатор A_f результатов работы алгоритма обнаружения объектов.

3.1.1 Построение обучающей выборки

Обучение классификатора на реальных данных оказывается затруднительным из-за отсутствия больших выборок размеченных данных видеонаблюдения с известными положениями людей и параметрами камеры. Поэтому используется синтетическая выборка, описанная в разделе 2.3.1. В качестве признакового описания результатов обнаружения выступают параметры 1) обнаруженного региона головы человека и 2) используемой камеры. Таким образом, входом алгоритма классификации является вектор, состоящий из:

- результата обнаружения $o = (x_o, y_o, s_o)$ объекта базовым алгоритмом A_b;
- позы $l_c = (h, t, r)$ камеры;
- фокусного расстояния f камеры.

Формально, с учетом особенностей построенной выборки задача классификации срабатываний детектора относится к классу задач поиска аномалий. Отличительной особенностью таких задач является отсутствие примеров отрицательного класса (ошибки детектора) в обучающей выборке. В своей работе я предлагаю способ моделирования объектов такого класса. Для их построения я объединяю две стратегии: 1) использование обнаружений, характерных для последовательностей обучающей выборки с отличающимися параметрами калибровки камеры; 2) использование регионов изображения с произвольным положением и размером. Первая стратегия позволяет обучать классификатор, дискриминативный к параметрам калибровки камеры. Вторая стратегия позволяет отличать корректные обнаружения от ложных срабатываний базового детектора A_b на фоновом изображнии. При построении отрицательных примеров обучающей выборки эти стратегии используются в отношении 1:9.

3.1.2 Построение классификатора

Для решение задачи классификации я использую полносвязную нейронную сеть, состоящую из 5 скрытых слоёв, с ReLu в качестве функции активации. Каждый скрытый слой имеет 20 нейронов. Количество скрытых слоев сети и нейронов в них выбиралось по валидационной выборке. Последний слой имеет 1 выход, к которому применяется логистическая функция. Выход нейронной сети интерпретируется как вероятность принадлежности к классу «правдоподобных» обнаружений.

3.2 Обучение и экспериментальная оценка

3.2.1 Обучение

Обучение классификатора производится с помощью метода Adam [54] с помощью логистической функции потерь. Обучение происходит на на синтетической выборке, состоящей из 24143 сцен, с помощью библиотеки caffe [57]. Скорость обучения понижалась каждые 1500 итераций со степенной скоростью $\gamma = 0.95$. В процессе обучения значение площади под кривой точности/полноты достигает значения 0.93 на валидационной выборке.

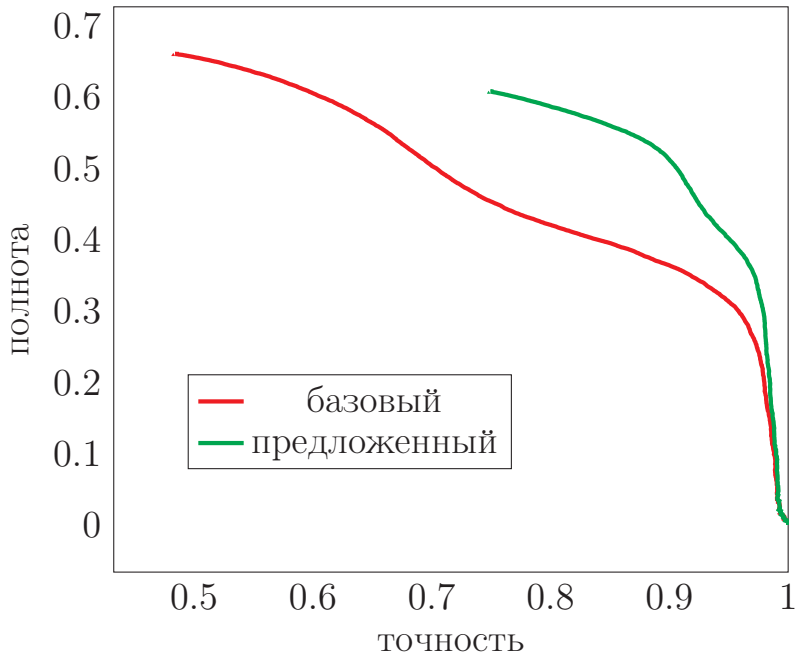


Рисунок 3.1 — Изменение качества обнаружения при применении обученного классификатора к реальным реальным данным видеонаблюдения.

3.2.2 Экспериментальная оценка на реальных данных

Для проведения экспертной оценки на реальных данных необходимо знать положение камеры для каждого кадра выборки. Это затрудняет использование стандартных баз размеченных изображений.

Поэтому при тестировании используется выборка TownCentre [26]. Алгоритм A_f классифицирует как «правдоподобные» обнаружения, на которых его уверенность превосходит значения 0.25. Обнаружение считается корректным, если его пересечение с регионом головы человека в разметке занимает не менее 25% их объединения. Сравнение качества исходного и полученного детектора (см. рис. 3.1) показывает, что предложенный алгоритм фильтрации позволяет увеличить точность на 18% при уменьшении полноты на 2.1%.

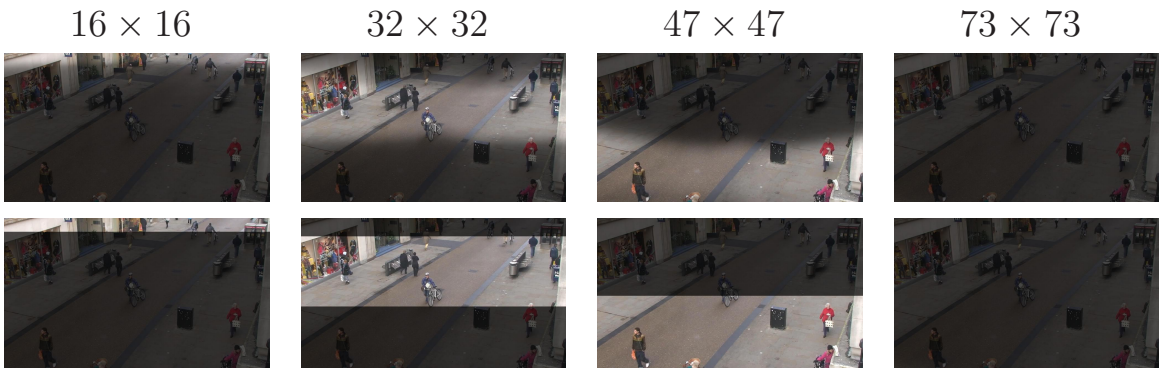


Рисунок 3.2 — Результаты применения классификатора обнаружений. Яркостью обозначено правдоподобие наблюдения головы заданного размера предсказанное классификатором (а) и используемое детектором (б).

3.2.3 Интеграция с алгоритмом детектирования

Предложенный алгоритм классификации A_f позволяет повысить производительность композиции $g(A_b, A_f)$ по сравнению с базовым детектором A_b. Он задает статические ограничения на допустимые положения изображения головы человека на кадре. На рисунке 3.2 (первая строка) показаны вероятности «допустимых» обнаружений разных размеров для камеры, соответствующей выборке TownCentre. Для каждого размера изображения головы человека классификатор предсказывает маску регионов, где располагаются «допустимые» обнаружения.

В случае статичной камеры положения описанных регионов не меняются с течением времени. Эта информация используется для увеличения производительности композиции $g(A_b, A_f)$ в сравнении с базовым алгоритмом обнаружения A_b. Если A_b использует пирамиду изображений, то для каждого её уровня обрабатывать необходимо лишь небольшой регион, предсказанный классификатором (рис 3.2). Для алгоритмов, производящих регрессию множества заранее заданных прямоугольников («якорей»), для каждого окна изображения классификатор предсказывает, какие «якори» не могут соответствовать правдоподобным положениям объектов интереса.

В диссертационной работе при тестировании в качестве базового детектора используется алгоритм, предложенный в [53]. При тестировании на выборке TownCentre обученный классификатор A_f указывает, что необходимо обработать лишь 21.44% окон на всех уровнях пирамиды. Использование произвольной маски обрабатываемых окон понижает эффективность обрабоки изображения. Поэтому в предложенной реализации алгоритм обрабатывает прямоугольные регионы изображения на каждом уровне пирамиды (см. рис. 3.2 вторая строка). Это соответствует обработке 24.03% всех окон. Такой подход позволяет увеличить производительность алгоритма локализации изображений голов людей с 20.03 до 34.36 кадров в секунду на выборке TownCentre.

3.3 Заключение

В главе был предложен алгоритм обнаружения людей на изображении, полученном камерой с известными параметрами калибровки. Предложенный алгоритм представляет суперпозицию $g(A_b, A_f)$ базового алгоритма обнаружения людей на изображении A_b и классификатора его результатов A_f. Использование суперпозиции позволяет выбрать в качестве базового любой алгоритм обнаружения людей на изображении. Построение классификатора A_f осуществляется с помощью машинного обучения на синтетической выборке видеонаблюдения. Экспериментальная оценка показала, что построенная суперпозиция позволяет повысить точность и скорость обнаружения людей на изображении.

Результаты главы были опубликованы в [58].

Глава 4. Сопровождение людей в видеопоследовательности

В данной главе рассматривается задача сопровождения людей в видеопоследовательности. Под сопровождением подразумевается построение траектории движения каждого человека, присутствующего в сцене. Можно считать, что задача сопровождения является расширением задачи обнаружения людей на изображении, где необходимо указать положение каждого человека не на одном, а на всех кадрах видео, где он присутствовал (рис. 4.1 вторая строка). Поэтому многие алгоритмы сопровождения людей в видео используют обнаружение человека на изображении в качестве первого этапа обработки.

Формально, входом алгоритмов сопровождения людей является видеопоследовательность $\{I_t\}_t$, а выходом множество траекторий $\{T_j\}$ движения людей. Каждая траектория T описывается последовательностью ограничивающих прямоугольников $\{b^t\}$ изображения одного человека в сегменте видео $t \in [t_b, t_e]$, где он присутствовал:

$$T = \left\{b^t\right\}_{t=t_b}^{t_e}$$

4.1 Базовый алгоритм

Как описано в разделе 1.3, для сопровождения людей используется подход сопровождения через обнаружение. Он заключается в последовательном решении задач обнаружения людей на кадрах видеопоследовательности и объединение результатов обнаружения в группы, соответствующие одному человеку, — траектории (рис. 4.1).

Среди множества алгоритмов сопровождения людей в видео в качестве базового подхода был выбран алгоритм [26]. Он выглядит наиболее перспективным для практического использования, поскольку позволяет строить

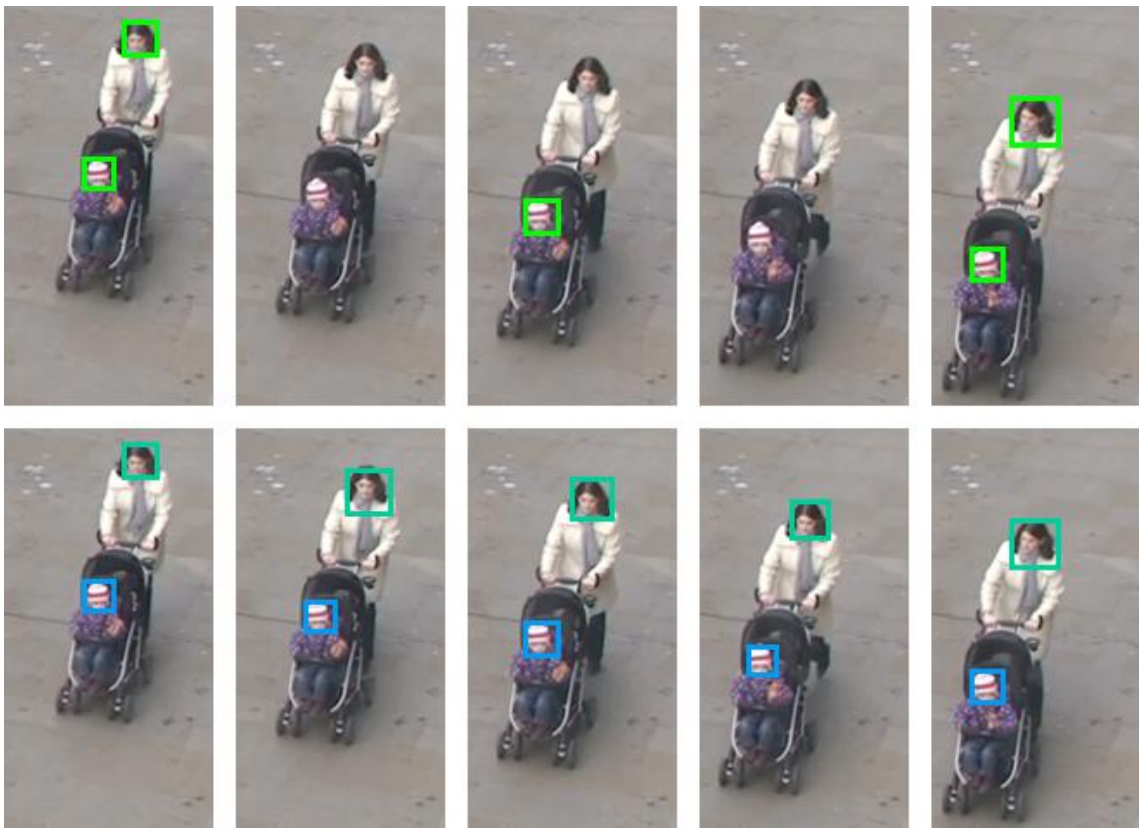


Рисунок 4.1 — Визуализация подхода сопровождения-через-обнаружение. В первой строке прямоугольниками показаны результаты верного обнаружения людей на разреженном множестве ключевых кадров: первом, третьем и пятом. Во второй строке показаны положения людей, полученные при сопровождении.

траектории людей в случае применения алгоритма обнаружения даже к разреженному множеству кадров (в дальнейшем называемых ключевыми), что существенно повышает скорость обработки данных.

Базовый алгоритм состоит из четырех этапов:

1. обнаружение изображений людей (их голов) на кадрах видеопоследовательности;
2. построение множества треклетов D — фрагментов траектории во временной окрестности кадра, где человек был обнаружен;
3. объединение треклетов в траектории;
4. восстановление положения людей на кадрах, где они не были обнаружены.

Задаче обнаружения людей на кадрах видеопоследовательности посвящена глава 3 данной диссертации. В следующих подразделах подробнее описываются остальные шаги решения задачи сопровождения базовым алгоритмом и предложенные модификации.

4.1.1 Построение треклетов

Для каждого обнаруженного изображения головы в момент времени t происходит построение треклета, содержащего информацию о положении человека во временной окрестности кадра I_t. На этом этапе описывается движение человека. В частности он позволяет определить направление его движения, что важно для дальнейшего построения траекторий.

Построение треклетов осуществляется с помощью визуального сопровождения обнаруженной области головы человека. В базовом алгоритме [26] предлагается использовать сопровождение нескольких ключевых точек [24] изображения головы алгоритмом KLT [59].

4.1.2 Объединение треклетов в траектории

На этом шаге решается задача разбиения множества треклетов D на группы, объединяющие треклеты одной траектории. Для упрощения изложения каждая такая группа треклетов также называется траекторией. Авторы базового алгоритма считают, что разные треклеты одного кадра не могут соответствовать одному человеку. Поэтому траектория не может содержать более одного треклета в момент времени t. Также каждый треклет соответствует некоторой траектории. Любое разбиение множества треклетов D на траектории $\{T_j\}$, удовлетворяющее этим ограничением, называется гипотезой H. Таким

образом, задача построения траекторий является задачей нахождения оптимальной гипотезы H^*.

Для описанной задачи вводится вероятностное распределение на множестве допустимых гипотез таким образом, что наиболее вероятная гипотеза представляет лучшее разбиение множества треклетов на траектории:

$$H^* = \arg\max_H p(H|D) = \arg\max_H p(H, D) = \arg\max_H p(H)p(D|H) \tag{4.1}$$

Таким образом, для описания алгоритма построения траекторий требуется:

– Задать вероятностное распределение на множестве гипотез $P(D|H)$;

– Определить метод поиска оптимальной гипотезы.

Модель движения

Модель движения задает функцию правдоподобия $P(D|H)$, описывающую разбиение треклетов на траектории. Важно отметить, что некоторые треклеты могут быть построены по ложным обнаружениям детектора, а соответственно не должны быть отнесены ни к какой траектории движения человека. Для решения этой проблемы в базовом методе каждая траектория содержит дискретную бинарную переменную — метку класса, принимающую одно из двух значений: c_{ped} — «человек», c_{fp} — «ложное обнаружение».

В работе [26] используется предположение, что люди двигаются в сцене независимо:

$$P(D|H) = p(T_1, T_2, ..., T_J|H) = \prod_{T_j \in H} p(T_j|c_j) \tag{4.2}$$

А правдоподобие каждой траектории факторизуется согласно динамической байесовской сети рис. 4.2:

$$p(T = \{d_0, d_1, d_2, \ldots, d_{n-1}, d_n\}|c) = p(d_0|c) \prod_{i=1}^{n} p(d_i|d_{i-1}, c), \tag{4.3}$$

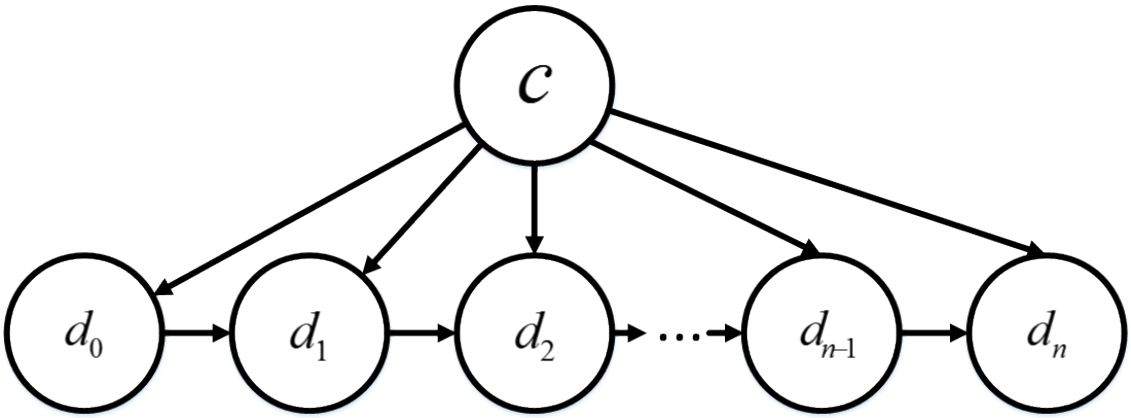


Рисунок 4.2 — Графическая модель, соответствующая траектории движения объекта. $\{d_i\}_{i=0}^{n}$ — множество треклетов рассматриваемой траектории, c — тип объекта: человек или ложное обнаружение.

где $p(d_0|c)$ — правдоподобие наблюдения первого треклета траектории и $p(d_i|d_{i-1}, c)$ — вероятность перехода от одного треклета к следующему. В базовом методе эти вероятности описывают такие характеристики треклета, как положение x обнаружения, его размер s, распределение скоростей движения m и оценки скорости Y движения человека в рамках треклета:

$$p(d_0|c) = p(s_0)p(x_0|c)p(m_0|c) \tag{4.4}$$

$$p(d_i|d_{i-1}, c) = p(s_i|s_{i-1})p(x_i|x_{i-1}, Y_{i-1}, c)p(m_i|c) \tag{4.5}$$

Для описания отличий предложенного метода от базового необходимо рассмотреть только факторы $p(x_0|c = c_{ped})$ и $p(x_i|x_{i-1}, Y_{i-1}, c = c_{ped})$, ограничивающие положение изображения человека на ключевых кадрах.

В базовом алгоритме используется предположение равномерного распределения положения человека в начале траектории вне зависимости от её типа:

$$p(x_0|c) \propto \frac{1}{\alpha}, \tag{4.6}$$

где α — размер изображения в пикселях.

Фактор $p(x_i|x_{i-1}, c = c_{ped})$ является наиболее значимым для поиска треклетов, относящихся к одной траектории. Он описывает изменение положения человека между кадрами и использует информацию о движении, содержащуюся в треклете. Этот фактор определяется как расстояние между обнаруженным

(а) (б)

Рисунок 4.3 — Визуализация фактора сходства положения треклетов траектории. Зелеными прямоугольниками показаны результаты работы детектора голов людей на ключевых кадрах, зеленой кривой — построенный треклет, красными прямоугольниками — предполагаемые положения человека на другом ключевом кадре. (а) $p(x_i|x_{i-1}, c = c_{ped})$ оценивает положение на следующем ключевом кадре; (б) $p(x_{i-1}|x_i, c = c_{ped})$ оценивает положение на предыдущем ключевом кадре.

положением изображения человека и предсказанными (рис. 4.3 (а)). Для предсказания положения изображения человека используются все оценки скорости его движения, т.е. средние значения скорости движения между кадром t' и ключевым кадром t:

$$p(x_i|x_{i-1}, Y_{i-1}, c_{ped}) = \frac{\alpha^{\delta_t}}{|Y_{i-1}|}\Sigma_{y\in Y_{i-1}} N(x_y^{i-1}, \Sigma_y + 2\Sigma_d) + (1 - \alpha^{\delta_t}) N(x_p^{i-1}, \Sigma_p + 2\Sigma_d), \tag{4.7}$$

где $|Y_{i-1}|$ — множество оценок скорости, полученных по треклету d_{i-1}. Подробно способ вычисления предсказанных положений x_y^{i-1} и x_p^{i-1} изображения человека на следующем ключевом кадре и степени доверия Σ_y, Σ_d, Σ_p предсказаниям описан в [26].

4.1.3 Алгоритм поиска оптимальной гипотезы

Для построения траекторий движения человека в сцене необходимо найти оптимальную гипотезу H^*:

$$H^* = \arg\max_H p(D|H)p(H) \tag{4.8}$$

В базовом алгоритме указано, что нет эффективного алгоритма для получения гипотезы, на которой достигается глобальный максимум функции распределения. Поэтому авторы предлагают использовать метод приближенного вывода — алгоритм MCMC DA. Данный алгоритм предполагает построение выборки гипотез из распределения $p(H, D)$ по схеме Марковских цепей. После этого гипотеза с наибольшей апостериорной вероятностью принимается в качестве приближения оптимальной гипотезы.

4.1.4 Восстановление положения

На последнем шаге базового алгоритма происходит определение положения каждого человека на каждом кадре видеопоследовательности. Для решения этой задачи используется линейная интерполяция положения человека между ключевыми кадрами, где он был обнаружен детектором.

Таблица 5 — Результаты сопровождения на выборке TownCentre

Результаты тестирования	MOTA	MOTP
Benfold и др. [26]	0.454	0.508
Кросс-корреляция шаблонов	0.507	0.524
«Стая точек»	0.519	0.525
«Стая точек» + регионы входа/выхода	0.54	0.522

4.2 Предложенный алгоритм

Для повышения точности построения траекторий предлагается новый алгоритм сопровождения, основанный на работе [26], но отличающийся от него по нескольким ключевым параметрам.

Качество построения траекторий зависит от точности определения, какие треклеты соответствуют одному человеку. В базовом алгоритме для решения этой задаче оценивается согласованность положения обнаруженных изображений человека одной траектории на основе факторов $p(d_0|c)$ и $p(d_i|d_{i-1}, c)$. Поэтому предложенные модификации направлены на повышение точности построения треклетов и изменение описанных факторов. В работе предлагаются несколько модификаций базового алгоритма:

1. для обнаружения голов людей на ключевых кадрах видеопоследовательности используется метод, предложенный в главе 3;
2. предлагается использование надежного метода визуального сопровождения для повышения точности построения треклетов;
3. предлагается использование согласованности положения не только последующего обнаружения, но и предыдущего по времени;
4. добавляется оценка регионов входа в сцену и априорное предпочтение на положение первого треклета траектории.

Рисунок 4.4 — Положения уголков (отмечены красными крестами), полученные при сопровождении в течение 75 кадров алгоритмом KLT (слева) и алгоритмом «Стая точек» (справа).

4.2.1 Построение треклетов

Точность локализации человека внутри треклета играет важную роль при объединение результатов обнаружения людей в траектории их движения. Можно выделить два типа ошибок при построении треклетов: 1) неточная локализация человека на кадре и 2) потеря сопровождаемого человека. Второй тип ошибок проявляется, когда происходит неверное предсказание положения человека на следующем ключевом кадре. Важно отметить, что этот тип ошибки наиболее критичен, поскольку может приводить к объединению треклетов, соответствующих разным людям, в одну траекторию. Решающую роль при предсказании положения человека на следующем ключевом кадре играют оценки скорости, соответствующие треклету.

Для построения треклетов в базовом методе применяется визуальное сопровождение алгоритмом KLT, реализующем отслеживание множества уголков на изображении головы человека. Этот алгоритм не накладывает ограничения на взаимное расположение сопровождаемых точек, что может приводить к ошибкам при построении треклета (рисунок 4.4). Поэтому для построения более надёжных оценок скорости предлагается использовать алгоритм «Стая

точек» [25]. В отличие от KLT алгоритм «Стая точек» позволяет обнаруживать и повторно инициализировать уголки, положение которых неверно определено на текущем кадре. Так же, когда изображение головы слабо текстурировано, на нем не удается найти достаточное количество уголков, которые можно было бы надежно сопровождать алгоритмами KLT и «Стая точек». Такая ситуация характерна, когда изображение головы имеет небольшие размеры, или человек повернут спиной к камере. В этом случае сопровождение на основе кросс-корреляции шаблонов [22] даёт более надёжные результаты.

4.2.2 Оценка согласованности положения человека

Согласованность результатов обнаружения людей, отнесенных к одному человеку, в базовой работе описывается фактором (4.7). Этот фактор учитывает только оценки скорости движения треклета d_{i-1}, построенного для предыдущего ключевого кадра $I_{t_{i-1}}$. Однако, было замечено, что использование оценок скорости не только треклета d_{i-1} , но и треклета d_i улучшает качество сопровождения. Поэтому, распределение положения треклета d_i предлагается определить следующим образом:

$$p(x_i|x_{i-1}, Y_i, Y_{i-1}, c_{ped}) = \beta p(x_i|x_{i-1}, Y_i, c_{ped}) + (1-\beta)p(x_i|x_{i-1}, Y_{i-1}, c_{ped}), \quad (4.9)$$

где Y_i и Y_{i-1} — множества используемых оценок скорости треклетов d_{i-1} и d_i соответственно. Фактор $p(x'|x, Y, c_{ped})$ определяется согласно (4.7). Параметр β указывает какой оценке дается большее предпочтение и зависит от длины треклетов d_i и d_{i-1}:

$$\beta = \frac{|Y_{i-1}| + 1}{|Y_{i-1}| + |Y_i| + 2} \quad (4.10)$$

В отличие от базового алгоритма, где предлагается использовать все оценки скорости треклета, в данной работе используются лишь те оценки скорости,

которые получены из положения сопровождаемого человека в момент обнаружения и его положений между ключевыми кардами, соответствующими треклетам d_i и d_{i-1}. Благодаря этому, во-первых, используются только самые надежные оценки скорости, во-вторых, предполагается равномерность движения человека только между соседними обнаружениями.

4.2.3 Ограничение положения первого обнаружения траектории

Для использования априорных предпочтений о положении человека на изображении необходима семантическая информация о наблюдаемой сцене: положение дорог, неба, домов и т.д. Извлечение такой семантической информации требует сложного анализа сцены и большого количества вычислительных ресурсов [60]. Поэтому для первого обнаружения человека в траектории предлагается использовать только близость к областям входа в сцену:

$$p_e(x_0) = N(\rho(x_0, enter)|0, \sigma_d^2), \tag{4.11}$$

где $\rho(x_0, enter)$ — расстояние от положения первого треклета x_0 до ближайшего региона входа в сцену. В результате априорное распределение первого треклета траектории представляется в виде смеси распределений (4.11) и (4.6):

$$\tilde{p}(x_0) = \frac{p_e(x_0) + p(x_0)}{2}. \tag{4.12}$$

Наиболее простым способом задания региона входа в сцену является выбор границы изображения в качестве неё. Однако, это может приводить к неверной работе алгоритма в случае, когда линия горизонта находится на изображении или в сцене присутствуют преграды, например, стена здания (рис. 4.5). Поэтому в работе предлагается определять регион входа в сцену при сопровождении.

В этой диссертационной работе предполагается, что данный регион образует выпуклый многоугольник на изображении (рисунок 4.5 (а)). Тогда

(а) (б)

Рисунок 4.5 — Поиск областей входа в сцену. (а) найденная область входа в сцену; (б) визуализация траекторий экспертной разметки

расстояние до области входа в сцену можно определить, как расстояние до границы этого многоугольника. В качестве такого многоугольника выбирается выпуклая оболочка положений первых и последних обнаружений человека, входящих в надежные траектории. Такое представление области входа в сцену позволяет проводить его обновление каждый раз после ассоциирования треклетов с траекториями.

Среди недостатков такого подхода можно выделить неустойчивость к выбросам. Например, если область входа в сцену была ошибочно увеличена, то она никогда не сможет быть уменьшена в дальнейшем. В связи с этим необходимо учитывать только надёжные траектории. В предложенном алгоритме надёжными считаются траектории, содержащие не менее 5 треклетов человека. Можно заметить, что построенная область входа в сцену соответствует границе области, содержащей траектории экспертной разметки (рис. 4.5 (б)).

Другим недостатком предложенного подхода является учёт препятствий на границе сцены в качестве регионов входа. Примером таких препятствий может служить стена здания изображенная на рисунке 4.5 (а).

4.3 Экспериментальная оценка

Экспериментальная оценка предложенного алгоритма проводилась на открытой базе TownCentre [26]. Она содержит видеопоследовательность высокого разрешения (1920x1080/25fps), снятую статичной камерой. Так же присутствует экспертная разметка, содержащая 71500 размеченных положений голов людей. В среднем на каждом кадре присутствуют 16 человек.

Для оценки качества сопровождения используются критерии MOTA и MOTP [61]. MOTA описывает качество построения траекторий людей, а MOTP — точность определения положения людей на кадрах видеопоследовательности. В связи с этим наиболее информативным является критерий MOTA. Значение критерия MOTA не превосходят 1, а значения критерия MOTP неотрицательны. Увеличение значения критерия MOTA свидетельствует об улучшении качества построенных траекторий людей, в то время как уменьшение значения MOTP говорит о повышении точности определения положения людей.

Результаты тестирования представлены в таблице 5. Алгоритмы, представленные во второй и третьей строках таблицы, различаются способом построения треклетов. Использование алгоритма «Стая точек» дает лучшие результаты на тестовой базе TownCentre, но качество сопровождения данным алгоритмом зависит от качества изображения головы человека. На слабо текстурированных изображениях не всегда удается найти достаточное количество уголков. В такой ситуации алгоритм сопровождения на основе кросс-корреляции шаблонов более устойчив к потере головы человека.

В последней строке представлены результаты работы предлагаемого алгоритма при использовании алгоритма «Стая точек» и учёте областей входа в сцену.

4.3.1 Анализ алгоритма

Как было указано выше, предлагаемый метод сопровождения представляет конвейер из применяемых последовательно алгоритмов. Поэтому важным является вопрос о том, какая часть данного конвейера является самой слабой частью метода. Одним из наиболее эффективных способов ответа на этот вопрос является метод, основанный на последовательном исключении первых этапов конвейера из алгоритма.

Данный метод заключается в последовательной замене результатов, получаемых на начальных этапах конвейера, на результаты экспертной разметки. Это позволяет оценить, какой вклад в общую ошибку дает каждый этап.

Результаты анализа представлены в таблице 6. В первом столбце указан этап конвейера, который был заменен экспертной разметкой. Исключение составляет только первая строка, где приведены результаты работы конвейера целиком. Во втором и третьем столбце указаны значения критерием MOTA и MOTP соответственно. Из результатов анализа следует, что улучшение алгоритма построения траекторий может дать наибольшее увеличение надежности сопровождения (около 0,28 с 0,629 до 0,916). Также анализ показал, что предположение о линейности движения людей между обнаружениями является очень грубым приближением, поскольку замена алгоритма восстановления положения людей на промежуточных кадрах на их положение в экспертной разметке увеличивает надежность сопровождения более, чем на 0,07. Также об этом свидетельствует повышение значения критерия MOTP при использовании траекторий, полученных из экспертной разметки. Данное повышение означает, что алгоритм построения траекторий совершает больше всего ошибок в случаях нелинейного движения.

Анализ показывает, что используемый алгоритм поиска людей позволяет проводить надежное сопровождение людей, но при этом точность локализации

Таблица 6 — Анализ качества работы предложенного алгоритма сопровождения

	MOTA	MOTP
Весь конвейер	0.54	0.522
Поиск людей	0.562	0.143
Построение треклетов	0.629	0.145
Построение траекторий	0.916	0.383
Восстановление положения	0.988	0.083

оказывается низкой. Стоит отдельно отметить почему не удается получить идеальных результатов даже в случае, если заменить последний этап конвейера на результаты экспертной разметки. В предлагаемом алгоритме поиск объектов осуществлялся 5 раз в секунду, то есть лишь на одном из пяти кадров. Также для положения сопровождаемых людей не проводилась экстраполяция вне сегмента между первым и последним обнаружениями. Это приводит к тому, что длительность сопровождения человека в экспертной разметке во многих случаях превышает длительность сопровождения предлагаемым алгоритмом. На точность локализации людей на изображении повлияло предположение о равенстве ширины и высоты прямоугольника, ограничивающего изображение головы человека. Вклад этого предположения отражен в значении критерия MOTP в последней строке.

4.4 Заключение

В главе был предложен алгоритм сопровождения множества людей в видеопоследовательности, точность построения траекторий которого превосходит базовый за счет использования надёжного метода оценки сходства обнаружений человека на разных кадрах. На повышение надежности оказали влияние

два предложенных фактора: фактор сходства положения обнаружений при обратном сопровождении и фактор, штрафующий разрыв траектории вдали от области входа в сцену.

Результаты главы были опубликованы в [62; 63].

Глава 5. Определение позы человека в видеопоследовательности

В этой главе рассматривается задача определения позы человека в видеопоследовательности. Как описано в обзоре 1.4, в настоящий момент поза человека на изображении определяется как положение K его суставов. Таким образом, позой человека в момент времени t является последовательность точек $P^t \in \left\{ \{p_i^t\}_{i=1}^K | p_i^t \in \mathbb{R}^2 \right\}$ на изображении.

В своей работе я рассматриваю алгоритм определения позы человека в качестве следующего этапа обработки видеоданных после сопровождения. Поэтому моя постановка задачи определения позы в видеопоследовательности имеет следующий вид:

Вход: – видеопоследовательность $I = \{I_t\}_{t=1}^N$;
– траектория движения человека $T = \{b^t\}_{t=1}^N$

Выход: поза человека на каждом кадре $P = \{P^t\}_{t=1}^N$.

Известная траектория движения человека ограничивает область изображения, на которой производится определение его позы.

5.1 Математическая модель наблюдаемых данных

Я рассматриваю определение позы человека в видео как задачу минимизации функции энергии $E(P, \Theta | I, T)$, где Θ — скрытые параметры модели. Параметр Θ может включать как скрытые параметры позы человека на одном кадре (например, параметр размера), так и глобальные параметры модели человека (цветовая модель). Можно считать, что функция энергии $E(P, \Theta | I, T)$ задаёт ненормированную функцию правдоподобия позы в видеопоследовательности в виде $\tilde{p}(P, \Theta | I, T) = \exp\left(-E(P, \Theta | I, T)\right)$. В дальнейшем для упрощения

выкладок я буду неявно предполагать зависимость функции энергии от исходной видеопоследовательности и положения человека в каждом кадре, т. е. $E(P, \Theta) = E(P, \Theta | I, T)$.

Используемая модель наблюдаемых данных является обобщением модели позы человека на изображении на случай видеопоследовательности. Для этого базовая модель расширяется предположением о зависимости позы человека на разных кадрах. Стандартный подход к обобщению базовой модели соответствует следующей функции энергии:

$$E(P, \Theta) = \sum_{t=1}^{T} E_I(P^t, \Theta) + \sum_{t=1}^{T-1} E_T(P^{t+1}, P^t, \Theta), \tag{5.1}$$

где $E_I(P^t, \Theta)$ — модель позы человека на кадре, а $E_T(P^{t+1}, P^t)$ — модель изменения позы между кадрами. Такую модель можно рассматривать как марковскую цепь первого порядка, где состояние в каждый момент времени является многомерной величиной и описывает позу человека. Она состоит из двух частей: 1) базовая модели позы человека на изображении $E_I(P^t, \Theta)$ и 2) модель движения её суставов $E_T(P^{t+1}, P^t, \Theta)$.

5.1.1 Модель позы человека на изображении

Согласно обзору существующих методов в научных работах представлены две основные модели позы человека: модель из набора деформируемых частей и регрессионная модель позы.

Несмотря на то, что регрессионная модель на момент написания данной диссертации позволяет добиться лучших результатов определения позы на изображении, её обобщение на случай видеопоследовательности затруднено. Построенное отображение входного изображения I_t в позу человека P^t на нем не позволяет использовать априорные сведения, полученные на соседних кадрах.

Поэтому в качестве базовой модели позы человека на изображении была выбрана модель из набора деформируемых частей. Соответствующая ей марковская сеть определяет ненормированную функцию правдоподобия $\tilde{p}(P^t, \Theta^t) = \exp(-E_I(P^t, \Theta))$. Это свойство позволяет интегрировать её как часть большей графической модели, используя информацию с предыдущих кадров для построения априорных ограничений.

Модель из набора частей описывается с помощью функции энергии $E_I(P^t, \Theta^t)$, минимум которой определяется в качестве позы человека на изображении I_t:

$$E_I(P^t, \Theta) = \sum_{i=1}^{K} \varphi_i(p_i^t, s^t) + \sum_{(i,j)\in E} \psi_{(i,j)}^s(p_i^t, p_j^t, s^t), \tag{5.2}$$

Унарный потенциал $\varphi_i(p_i^t, s^t)$ можно рассматривать, как отклик алгоритма обнаружения сустава человека на заданном масштабе изображения. Парный потенциал $\psi_{(i,j)}^s(p_i^t, p_j^t, s^t)$ задаётся в виде квадратичной формы, зависящей от смещения между суставами. Функция энергии зависит от параметра размера человека s^t на текущем изображении, и не зависит от остальных скрытых параметров модели.

На практике параметр положения p_i^t сустава является дискретной величиной, определённой на изображении с некоторым шагом. Параметр размера s^t также является дискретным и соответствует разным масштабам при обнаружении суставов.

5.1.2 Модель движения

Наиболее простым способом задания модели изменения позы является предположение о независимости движения суставов:

$$E_T(P^{t+1}, P^t, \Theta) = \sum_{i=1}^{K} \psi_i^t(p_i^{t+1}, p_i^t, \Theta) \tag{5.3}$$

Так как модели движения разных суставов схожи, то достаточно рассмотреть её лишь для одного из них. Для упрощения обозначений в данном подразделе я опускаю индекс рассматриваемого сустава. Например, p^t используется для обозначения состояния рассматриваемого сустава на кадре I_t.

Такое расширение модели позы человека на изображении на случай видеопоследовательности использовался также в предыдущих работах. Например, в работе [48] предлагалась модель движения, предполагающее слабое изменение позы человека между кадрами:

$$\psi^t(p^{t+1}, p^t, \Theta) = \frac{1}{2s^{t^2}}(p^{t+1} - p^t)^{T-1} \left(\Sigma_p^p\right)^{-1} (p^{t+1} - p^t)$$

Таким образом, оптимальное значение такой модели движения достигается при постоянстве позы человека в видео. Изменение позы при движении оказывается «допустимым шумом».

В этой работе я расширяю эту модель движения. Я использую линейную динамическую систему для описания движения суставов тела человека. Для этого скрытое состояние Θ модели расширяется характеристикой движения каждого сустава. В работе я рассматриваю линейную модель движения суставов, т. е. состояние каждого сустава описывается его положением p^t на кадре и мгновенной скоростью движения $v^t \in \mathbb{R}^2$. По аналогии с позой человека, я обозначаю скорость всех суставов в видеопоследовательности через V. Если обозначить через $h^t = [p^t, v^t]$ состояние рассматриваемого сустава позы человека на кадре t, то предложенная модель движения принимает вид:

$$\psi^t(p^{t+1}, p^t, \Theta) = \frac{1}{2s^{t^2}}(h^{t+1} - Ah^t)^T \Sigma_p^{-1} (h^{t+1} - Ah^t)$$

$$A = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.4}$$

Допустимое отклонение от линейной модели движения задается симметричной положительно определенной матрицей $\Sigma_p \in \mathbb{S}_+$. Для уменьшения количества параметров я рассматривал только диагональные матрицы Σ_p следующего вида:

$$\begin{aligned}
\Sigma_p &= \left[\begin{array}{c|c} \Sigma_p^p & \Theta \\ \hline \Theta & \Sigma_p^v \end{array}\right] \\
\Sigma_p^p &= \alpha_p^{-1} I_{2\times 2} \\
\Sigma_p^v &= \alpha_v^{-1} I_{2\times 2} \\
\alpha_p &> 0, \alpha_v > 0,
\end{aligned} \tag{5.5}$$

Матрица Σ_p^p описывает допустимое отклонение положения сустава p^{t+1} от его линейного предсказания $p^t + v^t$, а Σ_p^v — допустимое изменение скорости сустава между кадрами.

Без дополнительной регуляризации модель движения (5.4) допускает неправдоподобно большое значение скорости движения сустава, так как ограничивает только её изменение. Для решения этой проблемы был добавлен фактор, задающий априорное предпочтение на скорость движения суставов на первом кадре:

$$\begin{aligned}
\psi^0(v^1, \Theta) &= \frac{1}{2s^{1^2}} v^{1^T} \left(\Sigma_p^{v^1}\right)^{-1} v^1 \\
\Sigma_p^{v^1} &= \alpha_{v^1}^{-1} I_{2\times 2}
\end{aligned} \tag{5.6}$$

Также моя модель ограничивает неправдоподобное изменение размера человека между кадрами:

$$\eta^t(s^{t+1}, s^t) = \frac{1}{2}\left(\frac{s^{t+1} - s^t}{s^t \sigma_s}\right)^2 \tag{5.7}$$

Таким образом, предложенная модель движения имеет следующий вид:

$$\sum_{t=1}^{T-1} \Psi(P^{t+1}, P^t, \Theta) = \sum_{i=1}^{K}\left(\psi_i^0(v_i^1) + \sum_{t=1}^{T-1}\psi_i^t(h_i^{t+1}, h_i^t, \Theta)\right) + \sum_{t=1}^{T-1}\eta^t(s^{t+1}, s^t) \tag{5.8}$$

5.1.3 Частные случаи

Предложенная модель имеет два интересных частных случая, описывающих предыдущие работы по определению позы.

Рассмотрим случай, когда $\alpha_v \to +\infty$, $\alpha_{v^1} \to +\infty$ и $\sigma_s \to +\infty$. Этот случай описывает значительное увеличение энергии в случаях, когда значение скорости какого-либо сустава отлично от 0:

$$\begin{aligned} \lim_{\alpha_{v^1}\to+\infty} \arg\min_{v^1} \psi^0(v^1) &= 0 \\ \lim_{\alpha_v\to+\infty} \arg\min_{v^t} \psi^t(h_i^{t+1}, h_i^t, \Theta) &= 0 \\ \lim_{\sigma_s\to+\infty} \eta^t(s^{t+1},s^t) &= 0 \end{aligned} \tag{5.9}$$

То есть оптимальным является решение, где все параметры скорости равны 0. При этом условии модель движения суставов имеет вид:

$$\begin{aligned} \psi^t(h_i^{t+1}, h_i^t, \Theta)|_{v^t=v^{t+1}=0} &= \frac{1}{2s^{t2}}(p^{t+1}-p^t)^T \left(\Sigma_p^p\right)^{-1}(p^{t+1}-p^t) \\ \psi^0(v^1)|_{v^1=0} &= 0 \end{aligned} \tag{5.10}$$

То есть предложенная модель становится эквивалентной модели движения, описанной в работе [48].

Рассмотрим другой частный случай. Если ослабить ограничения на небольшое изменение скорости суставов между кадрами, то модель (5.8) описывает независимое определение позы человека на каждом кадре. Действительно

$$\begin{aligned} \lim_{\alpha_v\to0+0} \psi^t(h_i^{t+1}, h_i^t, \Theta) &= 0 \\ \lim_{\alpha_{v^1}\to0+0} \psi^0(v^1) &= 0 \\ \lim_{\sigma_s\to+\infty} \eta^t(s^{t+1},s^t) &= 0 \end{aligned} \tag{5.11}$$

То есть графическая модель распадается на независимые связные компоненты, соответствующие позе человека на каждом кадре.

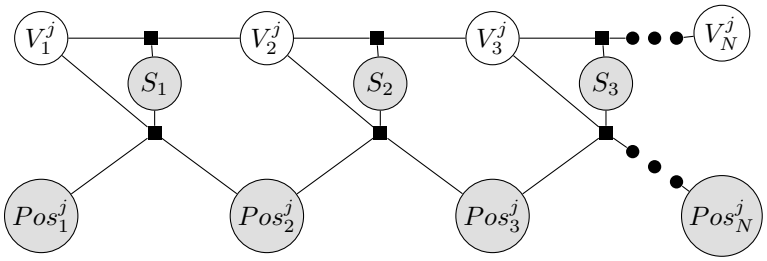


Рисунок 5.1 — Фактор граф, соответствующий задаче определения скорости. Наблюдаемые переменные отмечены серым

5.2 Метод оптимизации

5.2.1 Анализ модели

Для описания предложенного алгоритма оптимизации важно рассмотреть две задачи, связанные с представленной функции энергии:

1. определение скорости суставов при известных позе и росте человека человека в видео — $\arg\min_V E(P, \Theta)$;
2. определение позы и размера человека на кадре t при известных остальных параметрах модели — $\arg\min_{P^t, s^t} E(P, \Theta)$.

Определение скорости

Рассмотрим задачу определение скорости движения суставов при известных позе и росте человека в видео $V(P, \Theta_{\setminus V}) = \arg\min_V E(P, \Theta)$. Модель позы человека на изображении $E_I(P^t, \Theta)$ и модель изменения размера $\eta^t(s^{t+1}, s^t)$ не зависят от параметров скорости. Поэтому рассматриваемая задача эквивалентна задаче оптимизации:

$$V = \arg\min_V E(V|P, \Theta_{\setminus V}) = \arg\min_V \sum_{i=1}^{K} \left(\psi_i^0(v_i^1) + \sum_{t=1}^{T-1} \psi_i^t(h_i^{t+1}, h_i^t, \Theta) \right), \quad (5.12)$$

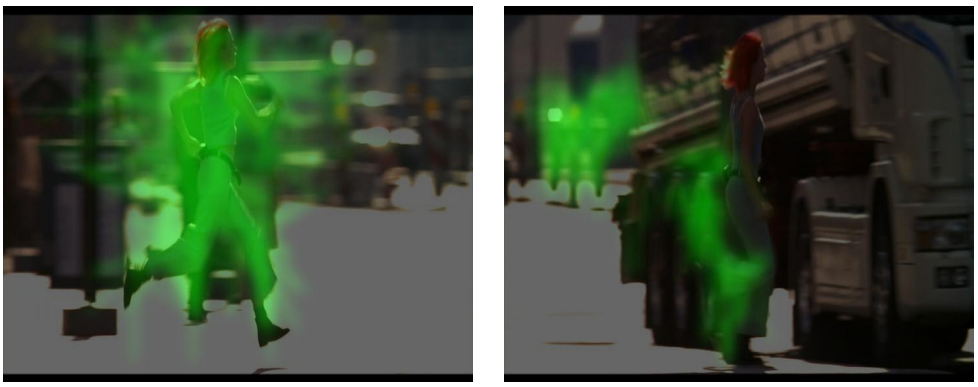
где $\Theta_{\setminus V}$ обозначает множество скрытых параметров, не содержащее параметры скорости V.

Из этой формы видно, что скорости разных суставов никогда не входят в одно слагаемое, а значит можно проводить оптимизацию скорости каждого сустава независимо:

$$\begin{aligned} V_i &= \arg\min_{V_i} \psi_i^0(v_i^1) + \sum_{t=1}^{T-1} \psi_i^t(h_i^{t+1}, h_i^t, \Theta) \\ V &= \cup_{i=1}^{K} V_i \end{aligned} \tag{5.13}$$

Рассмотрим поиск оптимального значения для скорости одного сустава. В дальнейшем для упрощения выкладок я буду опускать индекс текущего сустава. Функции энергии $E(V|P, \Theta_{\setminus V})$ соответствует фактор граф на рисунке 5.1. Можно заметить, что соответствующая графическая модель является марковской цепью первого порядка. Учитывая (5.4) и (5.5), оптимизируемую энергию $E(V|P, \Theta_{\setminus V})$ можно переписать в виде суммы унарных и парных потенциалов:

$$\begin{aligned} E(V|P, \Theta_{\setminus V}) &= \psi^0(v^1) + \sum_{t=1}^{T-1} \left(\psi^{t,u}(v^t|p^{t+1}, p^t, s^t) + \psi^{t,p}(v^{t+1}, v^t|s^t) \right) \\ \psi^{t,u}(v^t|p^{t+1}, p^t, s^t) &= \frac{(\Delta p^t - A^u v^t)^T \Sigma_p^{p-1} (\Delta p^t - A^u v^t)}{2{s^t}^2} \\ \psi^{t,p}(v^{t+1}, v^t|s^t) &= -\frac{(v^{t+1} - A^p v^t)^T \Sigma_p^{v-1} (v^{t+1} - A^p v^t)}{2{s^t}^2} \\ \Delta p^t &= p^{t+1} - A_p^u p^t \\ A &= \left[\begin{array}{c|c} A_p^u & A^u \\ \hline \Theta & A^p \end{array}\right] \quad A_p^u, A^u, A^p \in \mathbb{R}^{2\times 2} \end{aligned} \tag{5.14}$$



(а) (б)

Рисунок 5.2 — Визуализация наилучших гипотез позы человека на кадре. Интенсивность зелёной подсветки соответствует количеству построенных гипотез позы человека. На кадре (а) большинство обнаруженных гипотез близки к верной позе человека. На кадре (б) детектор не смог найти хорошее множество гипотез позы человека.

В данной постановке задача поиска минимума функции энергии $E(V|P,\Theta_{\setminus V})$ совпадает с задачей определения состояний линейной динамической системы (ЛДС):

$$\begin{aligned} V &= \arg\max_V p\left(V|\left\{\Delta p^t\right\}_{t=1}^{T-1},\left\{s^t\right\}_{t=1}^{T}\right) \\ \Delta p^t &\sim N(A^u v^t,(s^t)^2\Sigma_p^p) \\ v^{t+1} &\sim N(A^p v^t,(s^t)^2\Sigma_p^v) \\ v^1 &\sim N(\Theta,(s^t)^2\Sigma_p^{v^1}) \\ v^T &= A^p v^{T-1} \end{aligned} \tag{5.15}$$

Поиск наиболее вероятной конфигурации V осуществляется с помощью фильтра Калмана и РТС уравнений.

Определение позы и размера на кадре

Рассмотрим задачу определения позы P^t и размера s^t человека на некотором кадре при известных остальных параметрах модели $\arg\min_{P^t,s^t} E(P,\Theta)$:

$$E(P^t,s^t|P^{\setminus t},\Theta_{\setminus s^t}) = E(P,\Theta)\Big|_{P^{\setminus t},\Theta_{\setminus s^t}} = \\ = \left(\sum_{t'=1}^{T} E_I(P^{t'},\Theta) + \sum_{t'=1}^{T-1} E_T(P^{t'+1},P^{t'},\Theta)\right)\Big|_{P^{\setminus t},\Theta_{\setminus s^t}} \quad (5.16)$$

Модель позы человека в видео является марковской цепью первого порядка, поэтому рассматриваемая энергия имеет вид:

$$E(P^t,s^t|P^{\setminus t},\Theta_{\setminus s^t}) = E_I(P^t,\Theta) + \\ + \big(E_T(P^{t+1},P^t,\Theta) + E_T(P^t,P^{t-1},\Theta)\big)\Big|_{\substack{P^{t-1},P^{t+1},\\ s^{t-1},s^{t+1},\\ V^{t-1},V^{t+1}}} + C, \quad (5.17)$$

где C — константа, не зависящая от рассматриваемых параметров. Чтобы не рассматривать отдельно случаи граничные случаи, можно доопределить состояние модели в моменты времени $t=0$ и $t=T+1$ значениями в моменты времени $t=1$ и $t=T$ соответственно.

Учитывая (5.2) и (5.4), рассматриваемая энергия имеет вид:

$$E(P^t,s^t|P^{\setminus t},\Theta_{\setminus s^t}) = \sum_{i=1}^{K}\left(\underbrace{\varphi_i(p_i^t,s^t) + \psi_i^t(h_i^{t+1},h_i^t,\Theta) + \psi_i^t(h_i^t,h_i^{t-1},\Theta)}_{\varphi_i'(p_i^t,s^t)}\right) + \\ + \sum_{(i,j)\in E}\psi_{(i,j)}^s(p_i^t,p_j^t,s^t) + \underbrace{\eta^t(s^{t+1},s^t) + \eta^t(s^t,s^{t-1}) + C'}_{\varphi_s(s^t)} \quad (5.18)$$

Таким образом, условная модель человека $E(P^t,s^t|P^{\setminus t},\Theta_{\setminus s^t})$ на кадре I_t для каждого масштаба s^t представляет сумму унарных $\varphi_i'(p_i^t,s^t)$ и парных $\psi_{(i,j)}^s(p_i^t,p_j^t,s^t)$ потенциалов относительно положения суставов позы человека. Так как модель движения не добавляет парных потенциалов в энергию

$E(P^t, s^t | P^{\setminus t}, \Theta_{\setminus s^t})$, то эта энергия факторизуется согласно той же графической модели, что и модель позы $E_I(P^t, \Theta)$. Факторы $\eta^t(s^{t+1}, s^t)$ и $\eta^t(s^t, s^{t-1})$ задают апостериорное предпочтение на значение параметра размера позы.

Таким образом, условная модель позы человека $E(P^t, s^t | P^{\setminus t}, \Theta_{\setminus s^t})$ является обобщением модели позы человека на изображении $E_I(P^t, \Theta)$. Так как при оптимизации модели $E_I(P^t, \Theta)$ алгоритмы производят перебор по различным значениям параметра размера человека s^t, то для условной модели позы человека также выполняются свойства:

1. вычислительная сложность алгоритма поиска глобального оптимума равна $\mathcal{O}(KM)$, где K — количество суставов рассматриваемого скелета тела человека, M — количество допустимых положений одного сустава на изображении;
2. алгоритм построения множества N наилучших гипотез позы человека на изображении, отличающихся не менее чем на заданную величину равна имеет сложность $\mathcal{O}(NKM)$.

На рисунке 5.2 показан пример множества наилучших гипотез позы человека на изображении в ситуации, когда детектор верно локализует суставы и не может корректно определить позу человека.

Сложность поиска глобального оптимума

Каждой функции энергии, описывающей модель позы человека, соответствует марковская сеть. От её свойств зависит сложность алгоритмов поиска оптимального значения модели. В общем случае поиск глобального оптимума в графических моделях осуществляется алгоритмом распространения доверия. Его вычислительная сложность зависит от количества допустимых состояний минимальной группы вершин, разделяющих графическую модель на две части.

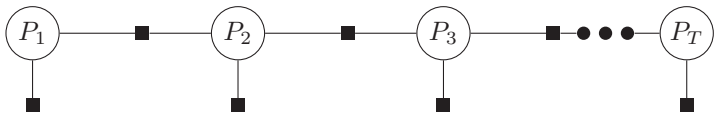


Рисунок 5.3 — Фактор-граф базовой модели позы.

Если графическая модель является деревом, то сложность алгоритма квадратично зависит от количества допустимых состояний для каждой вершины. Примером такой графической модели является модель позы человека на изображении при условии известного параметра масштаба. Специальный вид парных потенциалов, позволил сократить сложность вывода в графической модели до линейной зависимости от количества допустимых положений суставов на изображении.

Однако, при расширении этой модели, в марковской сети появляются циклы, а следовательно сложность поиска оптимального решения существенно возрастает. На рисунке 5.3 представлена графическая модель, соответствующая модели [48]. Её можно рассматривать как марковскую цепь первого порядка, где каждое состояние описывается множеством вершин одного кадра.

Рассмотрим вычислительную сложность алгоритма поиска глобального минимума для такой марковской цепи. Если на кадре имеется M возможных положений каждого сустав, то сложность алгоритма распространения доверия равна $\mathcal{O}(M^K T)$. Здесь существенно использован факт, что модель движения (5.8) является квадратичной формой, и для вычисления сообщений в сети может быть использован метод дистантного преобразования. Таким образом, сложность поиска оптимального множества поз линейно зависит от количества возможных поз человека на изображении. Так как значение параметра M может превосходить 10^4, а количество суставов исчисляться десятками, то алгоритм поиска точного решения оказывается не применим на практике. Уменьшив количество допустимых поз на кадре изображения, авторы [48] построили алгоритм поиска локального оптимума со сложностью $\mathcal{O}(N^2 T)$, где N — количество допустимых поз на одном кадре.

Предложенная в данной работе модель является расширением [48] и содержит её в качестве частного случая. Алгоритм распространения правдоподобия не может быть применён для поиска глобального оптимума предложенной модели, так как потребует слишком больших вычислительных ресурсов. Также скрытое состояние Θ дополнительно содержит непрерывные параметры скорости суставов, то есть описанная модель является дискретно-непрерывной. Это не позволяет использовать алгоритм [48] напрямую. Я предложил два алгоритма поиска оптимального значения построенной функции энергии. Первый алгоритм использует идею уменьшения количества допустимых гипотез позы человека на кадре, предложенную в работе [48], для поиска локального оптимума, а второй — метод построения выборки из распределения для уточнения результата.

5.2.2 Детерминированный алгоритм

Рассмотрим первый из предложенных алгоритмов. Его псевдокод представлен на листинге 1.

Идея алгоритма основана на последовательном решении двух задач: 1) построение гипотез позы человека на кадре согласно условной модели позы $E(P^t, s^t | P^{\setminus t}, \Theta_{\setminus s^t})$; 2) вычисление оптимального значения параметра скорости, согласно модели $E(V|P, \Theta_{\setminus V})$. Способ решения этих задач описан в разделе 5.2.1.

Для построения начальной инициализации я использую метод, предложенный в работе [48], то есть минимизируется энергия модели при условии, что скорость движения суставов равна нулю $E(P, \Theta)\Big|_{V=0}$. Для построенного приближенного решения позы и размера человека оценивается скорость движения суставов путём минимизации $E(V|P, \Theta_{\setminus V})$.

Data: I
Data: N
Result: P

$r \leftarrow \max(I.width(), I.height())$;
$P \leftarrow N\_best(E(P,\Theta)\Big|_{V=0}, N, r)$;
$V \leftarrow \arg\min_V E(V|P, \Theta_{\setminus V})$;
while $r > 1$ **do**
  $E_0 \leftarrow E(P,\Theta)$;
  for $t = \overline{1,T}$ **do**
    $\left\{\left(\overline{P}_k^t, \overline{s}_k^t\right)\right\}_k \leftarrow N\_best\left(E(P^t, s^t|P_{\setminus t}, \Theta_{\setminus s^t}), N, r\right)$;
    for $k = \overline{1,K}$ **do**
      $\overline{P}_k \leftarrow \left\{P_{\setminus t}, \overline{P_k^t}\right\}$
      $\overline{\Theta}_k \leftarrow \{\Theta, \overline{s}_k^t\}$
      $\overline{V}_k \leftarrow \arg\min_V E(V|\overline{P}_k, \overline{\Theta}_{k,\setminus V})$;
    end
    $(P,\Theta) \leftarrow \arg\min_{\overline{P}_k, \overline{\Theta}_k} \left\{E(\overline{P}_k, \overline{\Theta}_k)\right\}$;
  end
  if $E(P,\Theta) = E_0$ **then**
    $r \leftarrow \frac{r}{2}$;
  end
end

Алгоритм 1: Итеративный алгоритм построения позы человека в видео.

После этого итеративно до сходимости предложенный алгоритм выбирает произвольный кадр видеопоследовательности и улучшает оценку позы и скрытых параметров, ассоциированных с этим кадром. Согласно разделу 5.2.1, для произвольного кадра можно определить оптимальную позу и параметр размера на нём, если остальные параметры модели известны. Однако, для выбранного кадра I_t оценка скорости, полученная на предыдущем шаге может быть неоптимальной. Поэтому в предложенном алгоритме происходит построение набора из

N гипотез позы человека. Для каждой из них определяется значение параметров скорости, и выбирается гипотеза приводящая к наибольшему уменьшению функции энергии.

Метод построения гипотез позы человека на изображении, предложенный в работе [48], зависит от двух гиперпараметров: 1) радиуса r, определяющего минимальное различие между построенными гипотезами и 2) количества гипотез N, которое необходимо построить. Большое значение радиуса r позволяет получить существенно отличающиеся гипотезы позы и подходит для работы в случае неверно определённого значения скорости. Малые значения радиуса r позволяют производить локальный поиск позы в окрестности предыдущего решения. Поэтому в предложенном алгоритме этот параметр понижается с течением времени.

Поскольку функция энергии $E(P, \Theta)$ может содержать несколько локальных минимумов, алгоритмы локальной оптимизации не могут гарантировать нахождения оптимума. Детерминированный алгоритм 1 находит локальный минимум функции энергии $E(P, \Theta)$. Одной из ключевых проблем алгоритма является обновление позы человека кадр за кадром. Например, алгоритм не может покинуть локального минимума, где параметр размера позы человека, а следовательно и его поза, был неверно определен на некотором сегменте времени. Чтобы решить эту проблему был разработан стохастический алгоритм для уточнения результатов.

5.2.3 Стохастический алгоритм

Data: I
Data: $\tilde{P}$
Data: $\tilde{\Theta} = (\tilde{S}, \tilde{V})$
Data: N
Result: P

$H_0 \leftarrow (\tilde{P}, \tilde{S})$;
for $i = \overline{1,N}$ **do**
  $j \sim$ sample_step();
  $\overline{H} \leftarrow$ apply_step(H_{i-1}, j);
  $p_{Acc} \leftarrow \min\left(\frac{\tilde{p}(\overline{H})p_{Tr}(H|\overline{H})}{\tilde{p}(H)p_{Tr}(\overline{H}|H)}, 1\right)$;
  if $rand() < p_{Acc}$ **then**
    $H_i \leftarrow \overline{H}$;
  else
    $H_i \leftarrow H_{i-1}$;
  end
end
$(P, S) \leftarrow \arg\min_{H_i=(P_i,S_i)} E(H_i)$;

Алгоритм 2: Алгоритм сэмплирования для построения позы человека в видео.

Общее описание

Рассмотрим функцию $E_{\setminus V}(P, S) = \max_V E(P, \Theta)$, зависящую только от позы человека и параметров его размера. Способ вычисления указанного максимума описан в подразделе 5.2.1. Важно заметить, что функция $E_{\setminus V}(P, S)$

определена на дискретном множестве поз человека и допустимых значений размеров. Таким образом, эта функция задаёт ненормированную вероятность на множестве возможных поз людей в видео $\tilde{p}(P, S) = \exp(-E_{\setminus V}(P, S))$.

Предложенный алгоритм использует это представление для построения выборки из распределения $\tilde{p}(P, S)$ по схеме марковских цепей (MCMC). Построение выборки позволяет при длительном сэмплировании получить прецеденты из разных мод распределения $\tilde{p}(P, S)$. Таким образом, даже используя инициализацию из локального оптимума, предложенный алгоритм может найти лучшую гипотезу позы человека в видео. Гипотеза, обладающая наименьшим значением энергии, выбирается в качестве результата работы алгоритма.

Далее в рамках данного подраздела под гипотезой я понимаю текущее значение позы и параметров размера человека в видео:

$$H = (P, S)\,, \tag{5.19}$$

а соответствующее ей значение параметра скорости обозначаю через $V(H)$:

$$V(H) = \arg\max_{V} E(P, \Theta)\Big|_{(P,S)=H} \tag{5.20}$$

Существует множество способов построения выборки из распределения. Наиболее распространенные из них алгоритмы Гиббса и Метрополиса–Гастингса. Поскольку алгоритм Метрополиса–Гастингса позволяет обновлять группу параметров состояния модели (например, позу человека на нескольких кадрах одновременно), он был выбран для построения выборки. Высокоуровневое представление предложенного стохастического метода поиска оптимальной позы человека в видео представлено в алгоритме 2.

Для построения выборки алгоритмом Метрополиса–Гастингса необходимо также задать модель перехода $p_{Tr}(\overline{H}|H)$. Она описывает способ построения новой гипотезы позы $\overline{H}$ из предыдущей H. Согласно алгоритму Метрополиса–Гастингса новая гипотеза принимается с вероятностью $p_{Acc}(\overline{H}|H)$:

$$p_{Acc}(\overline{H}|H) = \min\left(\frac{\tilde{p}(\overline{H})p_{Tr}(H|\overline{H})}{\tilde{p}(H)p_{Tr}(\overline{H}|H)}, 1\right) \tag{5.21}$$

Важно отметить, что алгоритм Метрополиса–Гастингса не требует вычисление нормировочной константы распределения $\tilde{p}$. Более того для построения выборки достаточно иметь способ вычислить значение вероятности лишь отдельных гипотез. Это позволяет использовать метод сэмплирования для оптимизации также и более сложных моделей позы человека в видео, включающих глобальные параметры внешности такие, как цвет одежды и его комплекцию.

Модель перехода

Для построения выборки из распределения алгоритмом Метрополиса–Гастингса необходимо выбрать модель перехода $p_{Tr}(\overline{H}|H)$. Она описывает способ изменения положения суставов и значений параметра масштаба для построения новой гипотезы.

Выбор модели перехода является ключевым этапом при построении алгоритма оптимизации энергии $E_{\setminus V}(P, S)$. Наиболее простым вариантом модели перехода является равномерное распределение на множестве допустимых гипотез, не зависящее от текущего состояния H. Тогда, согласно (5.21), вероятность принять новую гипотезу пропорциональна отношению вероятностей этих гипотез в модели $\tilde{p}(H)$. В таком случае алгоритм Метрополиса–Гастингса отвергает большинство гипотез $\overline{H}$, если текущая гипотеза H соответствует локальному оптимуму.

Влияние данной проблемы уменьшается, когда модель перехода строит гипотезы со схожей или большей вероятностью. Поэтому предложенная модель перехода на каждом шаге случайно выбирает один из следующих типов изменения гипотезы:

1. случайное смещение суставов позы человека в момент времени t в локальной окрестности их текущего положения;

2. обновляет параметры позы и размера человека в момент времени t с учётом их значений в момент времени $t-1$;
3. использование одной из гипотез позы, согласно условной модели $E(P^t, s^t | P^{\setminus t}, \Theta_{\setminus s^t})$;
4. линейная интерполяция позы человека между моментами времени $[t_1,t_2]$.

Предложенная модель перехода между гипотезами обновляет позу человека либо на в некоторый момент времени t, либо на сегменте $[t_1, t_2]$. Для ускорения сходимости модель перехода чаще выбирает моменты времени t, где наблюдается наибольшее отклонение от предложенной модели $E_{\setminus V}(H)$, т.е. при выборе оптимального значения параметра скорости:

$$
\begin{aligned}
V(H) &= \arg\max_{V} E(P, \Theta)\Big|_{(P,S)=H} \\
\xi(t|H) &= \left(E_I(P^t, \Theta) + \frac{1}{2}\left(\Psi(P^t, P^{t-1}, \Theta) + \Psi(P^{t+1}, P^t, \Theta)\right)\right)\Big|_{\substack{(P,S)=H \\ V(H)}} \\
p(t) &\propto \max_{\tau} \xi(\tau) - \xi(t) \\
t &\sim p(t)
\end{aligned}
\tag{5.22}
$$

Выбор сегмента $[t_1, t_2]$ эквивалентен выбору двух граничных моментов времени t_1 и t_2.

Первые три типа перехода моделируют локальный поиск. Переход первого типа производит небольшое отклонение положения суставов P^t и размера позы s^t в выбранный момент времени. Шаг отклонения выбирается по следующему закону:

$$
\begin{aligned}
\overline{p}_j^{t\prime} &= p_j^t + \beta s^t; \\
s^{t\prime} &= s^t + \gamma \\
\beta &\sim N(0, \beta_p^{-1} I_{2\times 2}) \\
\gamma &\sim N(0, \gamma_p^{-1})
\end{aligned}
\tag{5.23}
$$

Второй тип перехода обновляет позу человека на кадре I_t её с учётом текущей гипотезы позы на предыдущем кадре, также добавляется небольшой

нормально распределённый шум согласно (5.23). Этот тип перехода эквивалентен минимизации модели движения между этими кадрами:

$$\left(\overline{P}^t, s^t\right) = \arg\min_{P^t, s^t} \left(\Psi(P^t, P^{t-1}, \Theta)\Big|_{\substack{(P^{t-1}, s^{t-1}) \in H \\ V(H)}} \right) \tag{5.24}$$

с последующим применением первого типа перехода к результату.

Переход третьего типа заключается в выборе случайной гипотезы позы, согласно модели $E(P^t, s^t | P^{\setminus t}, \Theta_{\setminus s^t})$. Алгоритм построения соответствующих гипотез описан в подразделе 5.2.1. Можно отметить, что применение только этого типа перехода и выбор наиболее правдоподобной гипотезы позы человека на кадре эквивалентно алгоритму 1. Однако для на данном шаге выбор гипотезы проводится стохастически, что позволяет исследовать разные моды распределения $\tilde{p}(P, S)$.

Переход четвёртого типа обновляет текущую гипотезу на интервале $[t_1, t_2]$. Для этого применяется линейная интерполяция положения суставов и параметра размера. Поскольку предложенная модель движения $E_T(P, \Theta)$ описывает прямолинейное равномерное движение суставов, то такое обновление гипотезы может уменьшить вклад парного потенциала в значение функции энергии.

Только переход первого типа является обратимым, то есть при построении выборки из распределения $\tilde{p}(H)$ вероятность вернуться в предыдущее состояние $p_{Tr}(H|\overline{H})$ должна быть равна нулю для других типов перехода. Однако, так как рассматриваемой задачей является поиск оптимальной позы человека в видео, то я использую предположение, что $p_{Tr}(H|\overline{H}) = p_{Tr}(\overline{H}|H)$ для перехода любого типа.

Предложенная модель настроена так, что вероятность выбирать первый тип изменения гипотезы в среднем 50 раз больше вероятности выбора остальных типов перехода. Таким образом, процесс работы алгоритма поиска оптимальной позы человека в видео можно разделить на два повторяющихся этапа:

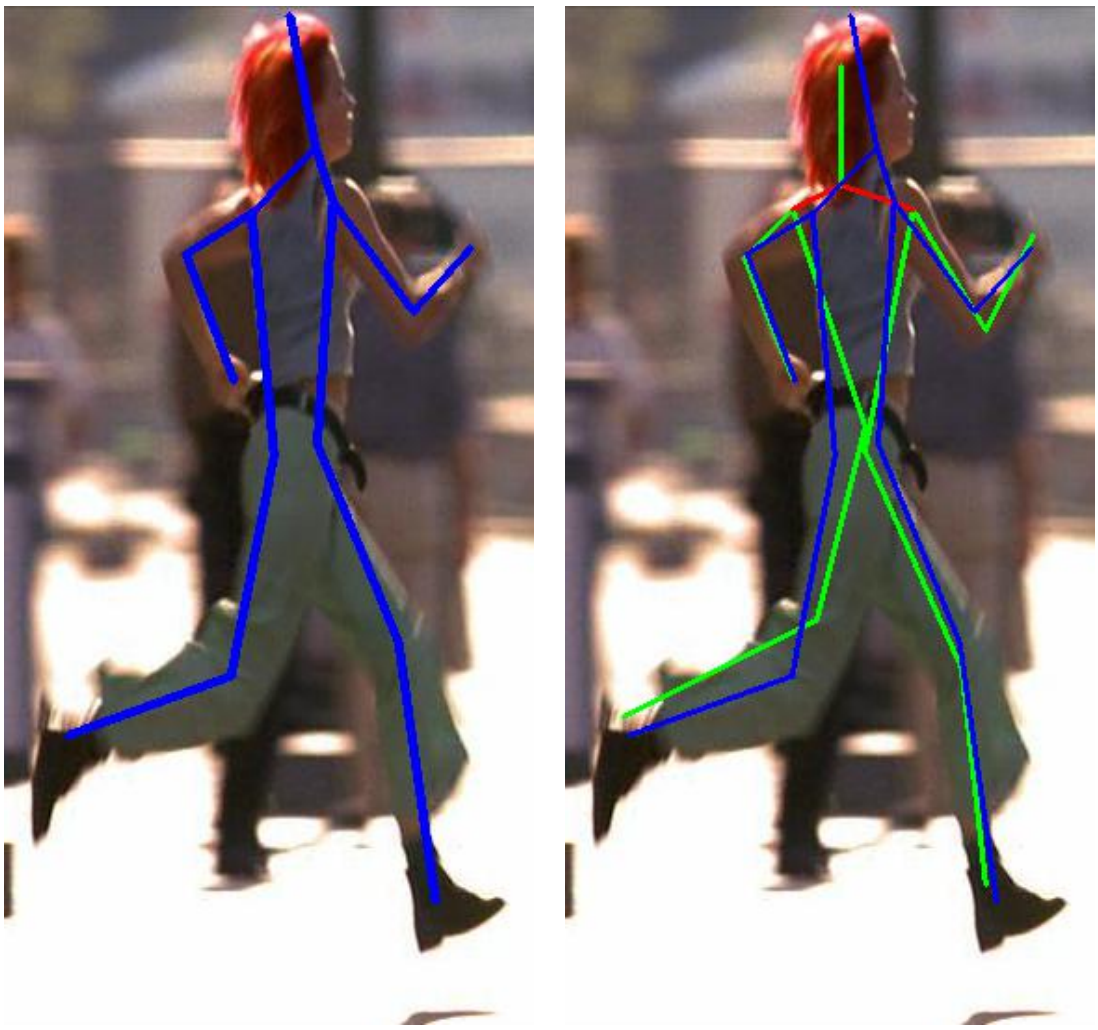
Рисунок 5.4 — Пример кадров тестовых последовательностей. Слева направо изображены кадры из последовательностей Walking, Pitching, Lola1, Lola2, соответственно.

1. построение выборки из распределения $\tilde{p}(H)$ (применение первого перехода типа);
2. выбор начальной гипотезы для сэмплирования (применение другого типа перехода);

5.3 Экспериментальная оценка

5.3.1 Выборка

Численная оценка предложенного алгоритма с базовым проводилась на выборке, предложенной в [48]. Примеры кадров этой выборки представлены на рисунке 5.4. Выборка состоит из 4 видеопоследовательностей, названных **Pitching**, **Lola1**, **Lola2** и **Walking**, различающихся по сложности. Видеопоследовательности **Pitching**, **Lola1** и **Lola2** содержат движение камеры. В видеопоследовательности **Pitching** присутствует изменение фокусного расстояния. Большинство последовательностей содержат одного человека в кадре, но в **Lola2** некоторые кадры содержат нескольких людей.



(а) (б)

Рисунок 5.5 — Визуализация позы как набор частей (отрезков). (а) экспертная разметка позы на кадре, (б) результат работы предложенного метода. Экспертная разметка показана синим. Зеленные части верно размечены, согласно PCP критерию, красные — ошибочно.

5.3.2 Результаты сравнения

Сравнение алгоритмов проводилось по критерию доли верно определённых частей (PCP), предложенному в [64]. Однако данный критерий имеет важный недостаток. Он интерпретирует позу как набор частей или отрезков (рис. 5.5 а) и оценивает корректность определения каждого из них независимо. Положение части считается верно определённой, если расстояние от концов отрезка до их корректных положений не отличается более чем на половину

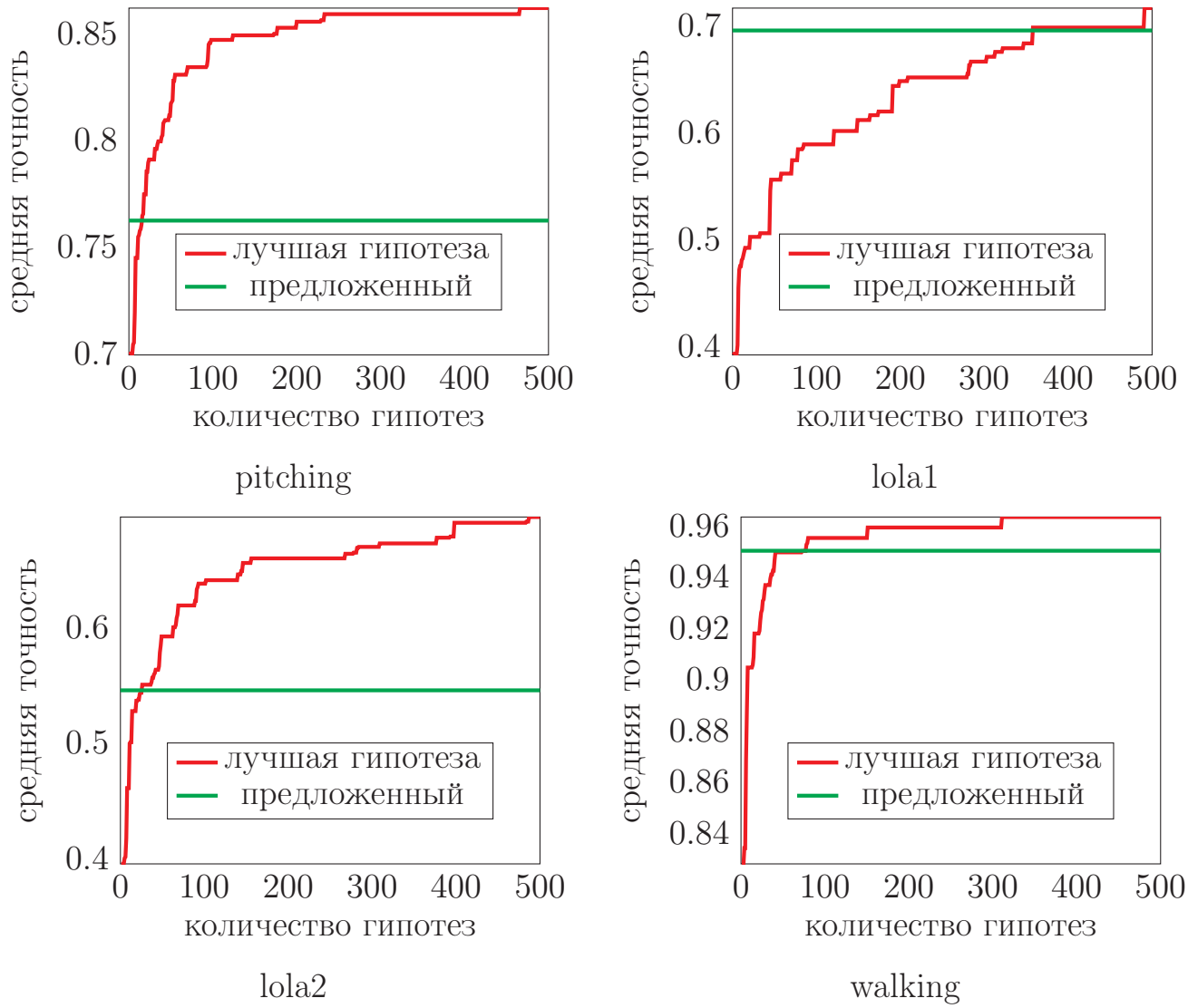


Рисунок 5.6 — Верхняя оценка (красная кривая) качества работы алгоритма [48] как функция от количества гипотез на кадре. Результат работы предложенного алгоритма показан зеленым.

длины отрезка в экспертной разметке. Так как размеры частей различны, то положение сустава может считаться верно определенным в составе одной части, и не верно в составе другой (рис. 5.5 б). Однако, поскольку PCP критерий до сих пор используется для оценки качества работы методов, я использую его для сравнения.

Для честного сравнения я не использую результаты сопровождения при инициализации, т. е. в начальный момент времени положение позы человека на изображении считается равномерным. Также параметры алгоритма α_p, α_v и σ_s были фиксированы для всей последовательностей выборки. Результаты

сравнения представлены в таблице 7. В качестве базового метода для сравнения был использован [48].

Таблица 7 — Результаты сравнения предложенного и базового метода. Сравнение производилось по среднему количествую верно локализованных частей (PCP). Результаты базового метода взяты из [48].

Алгоритм	walking	pitching	lola1	lola2
базовый	**0.950**	**0.797**	0.670	0.500
предложенный	**0.950**	0.762	**0.695**	**0.545**

Предложенный метод превосходит базовый на наиболее сложных последовательностях **Lola1** и **Lola2**. Алгоритм смог разрешить неопределенность с наличием нескольких людей в кадре на **Lola2** за счет использования информации о направлении и скорости движения.

Наиболее простой для алгоритмов определения позы человека в видео является видеопоследовательность **Walking**. В ней представлен один человек, идущий равномерно. Результаты базового и предложенного алгоритмов не отличаются на данной последовательности и связаны с ограничениями используемой модели позы человека на кадре [37].

При тестировании на последовательности **Pitching** предложенный метод показал более низкое значение критерия PCP в сравнении с базовым. Этот пример оказывается наиболее сложным для предложенного алгоритма, так как шаблон движения представленного спортсмена не соответствует линейной модели движения суставов.

Для оценки допустимых возможностей используемой модели человека [48] я оценил зависимость верхней границы качества построенных гипотез позы человека от их количества (рис. 5.6). Для оценки верхней границы на каждом кадре среди построенных гипотез выбиралась наилучшая по PCP критерию. Результаты показали, что предложенный подход выбирает позы близкие к оптимальным на последовательностях **lola1** и **walking**. На последовательностях **pitching** и **lola2** полученное решение существенно отличается от построенной верхней границы из-за сложных шаблонов движения конечностей человека и

движения камеры. Последовательности **walking** и **lola1** содержат тип движения, соответствующий сценарию видеонаблюдения.

Многие тестовые последовательности, рассматриваемые в работе [48], являются сложными для разработанного алгоритма, так как из-за движения камеры и изменения масштаба съемки нарушается предположение о равномерности движения суставов тела человека на изображении. Поэтому также была проведена оценка точности определения позы людей в сценарии видеонаблюдения со статичной камерой. С помощью автоматизированного средства, описанного в главе 6, была размечена подвыборка последовательности TownCentre [26]. Полученная разметка содержит 2000 поз людей в видео. Также суставы скелета человека, которые не видны на изображении, в разметке помечены, как невидимые. При оценке качества алгоритма сравнение производилось только по подмножеству частей тела, для которых оба сустава отмеченны, как видимые в экспертной разметке. На построенной выборке предложенный алгоритм показал качество 0.673 по метрике PCP, в то время, как базовый – 0.586. Повышение качества определения позы удалось добиться за счет использования информации о скорости движения суставов скелета человека. Действительно, в рассматриваемом сценарии движение большинства людей близко к равномерному. Поэтому предложенная модель позволяет лучше предсказать положение суставов тела человека при переходе к следующему кадру и восстанавливать положение суставов тела человека даже в случае их перекрытия.

5.4 Заключение

В главе был предложен алгоритм оценки позы человека в видеопоследовательности, учитывающий одновременно положение и скорость движения каждого сустава тела человека на кадре. Было показано, что предложенные ранее алгоритмы оценки позы, основанные на модели из набора частей, являются

частными случаями предложенной. За счёт использования информации о скорости движения суставов предложенный алгоритм позволил улучшить точность определения позы в сценарии с неподвижной камерой.

Результаты главы были опубликованы в [65].

Глава 6. Программная реализация

6.1 Общее описание

Предложенные в главах 2, 3, 4 и 5 алгоритмы были реализованы в виде отдельных модулей (рис. 6.1). На их основе, мной были разработаны и реализованы два программных средства. Первое решает задачу автоматического выделения и сопровождения всех пешеходов и изменений их позы в видеопоследовательности. Второе программное средство предоставляет автоматизированный инструмент для определения позы человека в видеопоследовательности.

6.2 Сопровождение людей и определение их позы в видео

Предложенные в главах 2, 3, 4 и 5 алгоритмы являются основой разработанного программного средства сопровождения людей и определения их позы в видео. На рисунке 6.2 схематично изображена блок-схема взаимодействия компонент разработанного приложения.

Входными данными является последовательность кадров, а выходом множество поз людей с указанием идентификатора траектории для каждой из них. Я предполагаю, что входная последовательность кадров имеет частоту не менее 25 кадров в секунду.

На первом этапе изображение обрабатывается детектором голов людей. У используемого детектора есть два режима работы. В режиме инициализации обрабатываются первые кадры видеопоследовательности. При этом положение камеры считается неизвестным и детектор совпадает с базовым, то есть кадр обрабатывается целиком на каждом уровне пирамиды. После применения детектора к 20 кадрам оцененное положение камеры считается надежным,

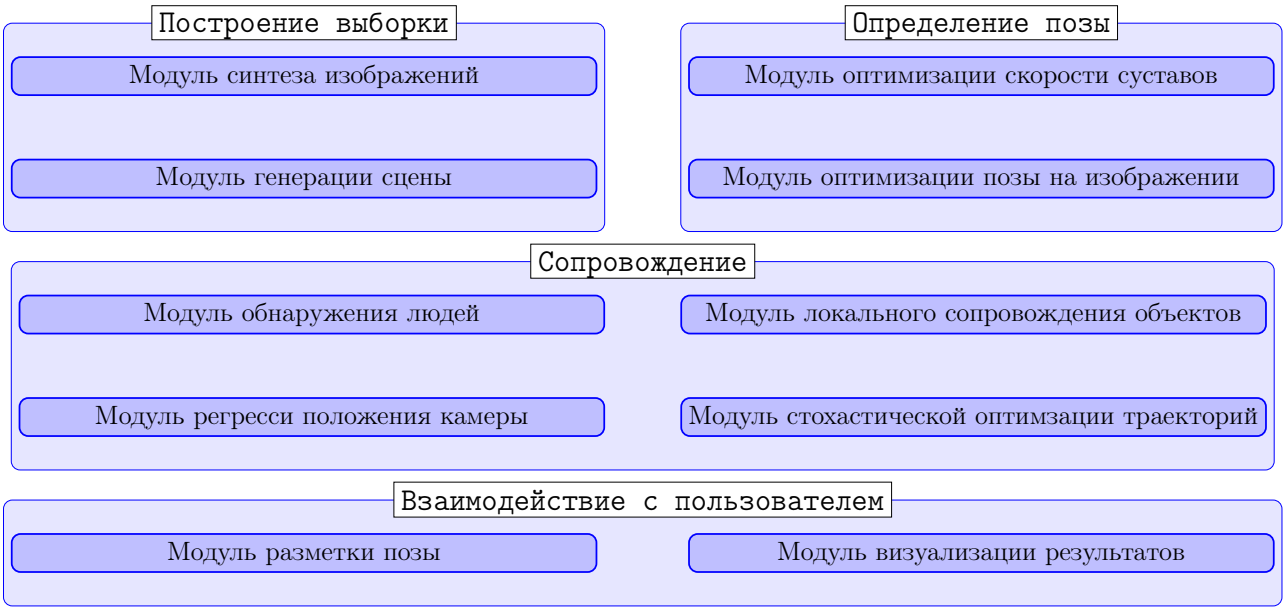


Рисунок 6.1 — Модули, разработанные и реализованные при выполнении диссертационной работы

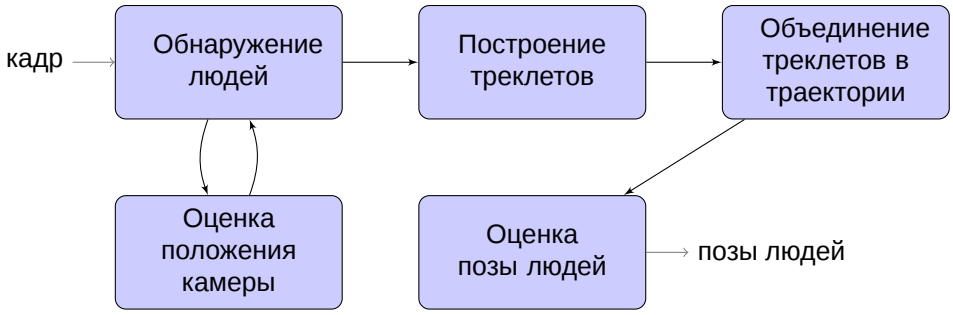


Рисунок 6.2 — Схема взаимодействия компонент, образующих программное средство сопровождения людей и определения их позы

и используется алгоритм, предложенный в главе 3. Для ускорения обработки данных детектор применяется к каждому 5-ому кадру, т.е. 5 раз в секунду. Однако, некорректно определенное положение камеры не позволит найти людей во входной видеопоследовательности. Поэтому раз в секунду к видеопоследовательности применяется базовый детектор, результаты которого используются для уточнения положения камеры. Результаты обнаружения и кадры видеопоследовательности передаются на следующий этап для построения треклетов.

Построение треклетов и объединение треклетов в траектории происходит, как описано в главе 4. Построение траекторий осуществляется по временному окну, содержащему треклеты за последние 200 кадров (8 секунд), а метки (идентификаторы траекторий) для обнаружений выдаются для центрального кадра этого временного окна. При построении траекторий выполняется 10000 итераций стохастической оптимизации для каждого кадра.

Модуль определения позы человека в видеопоследовательности получает кадр видеопоследовательности и результат сопровождения на кадре. Поскольку получаемый результат сопровождения соответствует кадру, который обрабатывался 4 секунды назад, в модуле предусмотрен буфер последних полученных кадров. Для траектории каждого человека из обрабатываемого кадра вырезается его изображение и выполняется поиск позы на всем видеофрагменте, его содержащем. Результатом работы модуля является поза человека на обрабатываемом кадре и идентификатор соответствующей траектории.

При реализации модулей использовался язык `C++`. Для ускорения работы детектора голов людей была написана реализация на `CUDA C++` с использованием `CUDA stream` для конвейерной обработки последовательных изображений. Для реализации матричных операций в детекторе позы использовалась библиотека `eigen`.

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.

6.3 Автоматизация построения экспертной разметки позы человека

В главе 5 описан предложенный алгоритм автоматического определения позы человека в видеопоследовательности. Однако, как показано в результатах, он не позволяет точно определить положение суставов на каждом кадре. Для развития алгоритмов определения позы человека в видео необходимо решить задачу автоматизации построения выборок видеопоследовательностей, содержащих разметку позы человека на каждом кадре. Для решения этой задачи было построено автоматизированное программное средство разметки позы человека на каждом кадре, основанное на предложенном в главе 5 алгоритме определения позы.

Предложенное средство позволяет пользователю указать дополнительную информацию двух типов:

1. виден ли сустав используемого скелета на текущем изображении
2. положение прямоугольника, ограничивающего положение сустава на текущем кадре

Эта дополнительная информация впоследствии используется в качестве дополнительных ограничений на множество допустимых поз человека на кадрах. Информацию первого типа можно описать с помощью бинарной переменной m_i^t, значение которой равно 1 только, когда i-ый сустав виден на изображении I_t. Информация второго типа описывается положением прямоугольника b_i^t на кадре I_t, внутри которого находится сустав. Как и в предыдущей главе, M^t и B^t обозначают векторы ограничений видимости и положения всех суставов позы человека на кадре I_t. В дальнейшем я считаю, что суставы, для которых пользователь не указал дополнительную информацию, видны, и их ограничивающий прямоугольник совпадает со всем изображением. Ниже будет показано, как осуществляется поиск позы человека в видеопоследовательности с учетом предложенных ограничений.

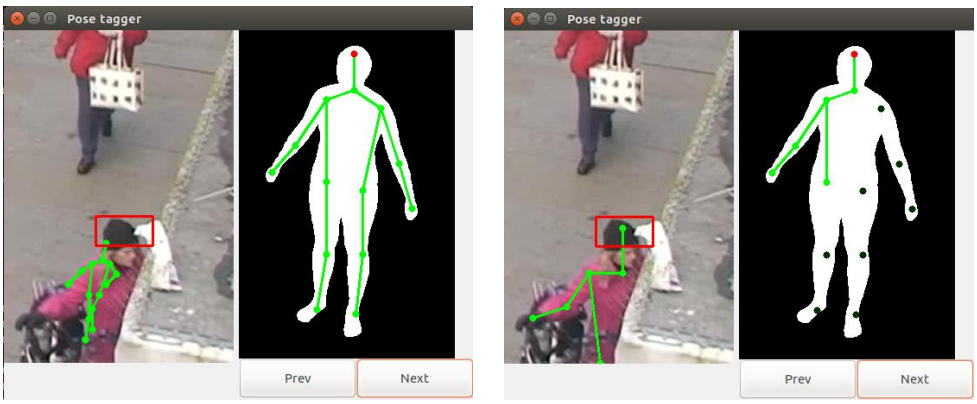
Решение задачи определения позы в видеопоследовательности сводится к безусловной оптимизации функции (5.1). С учетом дополнительной информации задача оптимизации приобретает вид:

$$\sum_{t=1}^{T} E_I(P^t, \Theta, M^t) + \sum_{t=1}^{T-1} E_T(P^{t+1}, P^t, \Theta) \to \min_{p_i^t \in b_i^t} \tag{6.1}$$

Важно отметить, что вносимые ограничения не влияют на используемую модель движения E_T. Ниже я укажу, как описанные ограничения изменяют модель оптимизируемую функцию и способ поиска оптимума.

Информация о том, виден ли сустав на кадре изображения влияет только на компоненту E_I оптимизируемой функции. Как было описано выше, эта компонента является суммой из двух факторов: $\varphi_i(p_i^t, s_i^t)$ — отклика детектора сустава на изображении и $\psi^s_{(i,j)}(p_i^t, p_j^t, s^t)$ — ограничения на взаимное расположение суставов. Если сустав i не виден на обрабатываемом кадре, то унарный потенциал $\varphi_i(p_i^t, s_i^t)$ в формуле (5.2) не может содержать информацию о положении этого сустава. Более того использование этого фактора может привести к некорректному определению позы человека, так как будет привносить дополнительный шум в оптимизируемую функцию. Поэтому я использовал фактор $m_i^t \varphi_i(p_i^t, s_i^t)$, зрачение которого равно исходному, если сустав помечен как видимый, и 0, иначае. Это позволяет не подстраивать позу человека под отклики детектора на изображении, и выбрать положение сустава только на основе положения соседних суставов.

Ограничение на положение сустава внутри заданного прямоугольника b_i^t требует использование методов оптимизации с ограничениями для определения позы человека. Однако используя метод неопределенных множетелей Лагранжа



(а) (б)

Рисунок 6.3 — Примеры пользовательского интерфейса и определенной позы человека (а) до и (б) после указания видимых на изображении суставов.

оптимизационную задачу (6.1) можно переписать в эквивалентном виде:

$$\sum_{t=1}^{T}\left(E_I(P^t,\Theta,M^t)+E_L(P^t,B^t)\right)+\sum_{t=1}^{T-1}E_T(P^{t+1},P^t,\Theta)\to min_{P\text{ s.t. }p_i^t\in b_i^t}$$
$$E_L(P^t,B^t)=\sum_i\delta(p_i^t,b_i^t) \tag{6.2}$$
$$\delta(p_i^t,b_i^t)=\begin{cases}0, & \text{если } p_i^t\in b_i^t\\ +\infty, & \text{иначе}\end{cases}$$

Важно отметить, что предложенный дополнительный фактор $\delta(p_i^t,b_i^t)$ является унарным. Таким образом, модель с учетом дополнительных ограничений, накладываемых пользователем, эквивалентна базовой, где вместо унарного фактора положения сустава $\varphi_i(p_i^t,s^t)$ используется фактор $\varphi_i'(p_i^t,s^t,m_i^t,b_i^t)=m_i^t\varphi_i(p_i^t,s_t)+\delta(p_i^t,b_i^t)$. Алгоритмы поиска минимума энекрии $E(P^t,\Theta)$ не зависит от вида унарных потенциалов. Поэтому они без изменений применяются для поиска оптимальной позы человека в видео с учетом внесенных ограничений.

Средство разметки было реализовано на языке `python2`. Разработанный алгоритм определения позы человека на изображении подключается в качестве стороннего модуля с использованием библиотеки `boost::python`. Для взаимодействия с пользователем используется библиотека построения графического интерфейса wxWidgets посредством модуля `wxPython`.

На рисунке 6.3 изображен пользовательский интерфейс приложения. В правой части экрана показана модель человека и расположение суставов на нем. Используя левую кнопку мыши на изображении сустава пользователь может указать значение параметра его видимости. С помощью правой кнопки мыши пользователь может выбрать другой сустав (отмечен красным на модели) и выделить область изображения, где он может находиться. На рисунке 6.3 (а) показан результат работы алгоритмы определения позы, когда указано допустимое положение одного сустава и все суставы считаются видимыми. На рисунке 6.3 (б) показан результат работы после того, как пользователь указал суставы отсутствующие на изображении. Кнопки пользовательского интерфейса позволяют переходить вперед и назад в рамках траектории указанного человека.

После внесения дополнительной информации пользователем алгоритм обновляет позу человека на всех кадрах видеопоследовательности, используя предыдущее решения в качестве инициализации.

Предложенное решение имеет несколько преимуществ в сравнении с полностью ручной разметкой позы человека на каждом кадре:

- чтобы исправить ошибки алгоритма, во многих случаях пользователю достаточно отметить допустимые регионы лишь для небольшого множества суставов скелета человека на кадре;
- в случае если поза человека была неверно определена на некотором сегменте видео, пользователю в некоторых случаях достаточно исправить результат лишь на одном кадре, а не размечать весь сегмент;
- при ручной разметке кадров последовательности независимо друг от друга возникает эффект случайного «дрожания» суставов между кадрами, связанный не с движением размечаемого человека, а с неточностями разметки. Предложенный метод разметки уменьшает этот эффект за счет использования фактора сглаживания E_T.

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.

Заключение

В ходе диссертационного исследования были получены следующие основные результаты:

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.

Список литературы

1. *Wang X.* Intelligent multi-camera video surveillance: A review // Pattern recognition letters. — 2013. — Т. 34, № 1. — С. 3—19.
2. *Caprile B.*, *Torre V.* Using vanishing points for camera calibration // International journal of computer vision. — 1990. — Т. 4, № 2. — С. 127—139.
3. Simultaneous vanishing point detection and camera calibration from single images / B. Li [и др.] // International Symposium on Visual Computing. — Springer. 2010. — С. 151—160.
4. *Liu J.*, *Collins R. T.*, *Liu Y.* Surveillance camera autocalibration based on pedestrian height distributions // British Machine Vision Conference (BMVC). — 2011.
5. Accurate self-calibration of two cameras by observations of a moving person on a ground plane / T. Chen [и др.] // Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on. — IEEE. 2007. — С. 129—134.
6. *Pflugfelder R.*, *Bischof H.* People tracking across two distant self-calibrated cameras // Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on. — IEEE. 2007. — С. 393—398.
7. Automatic inference of geometric camera parameters and inter-camera topology in uncalibrated disjoint surveillance cameras / R. J. den Hollander [и др.] // SPIE Security+ Defence. — International Society for Optics, Photonics. 2015. — С. 96520D—96520D.
8. Ptz camera network calibration from moving people in sports broadcasts / J. Puwein [и др.] // Applications of Computer Vision (WACV), 2012 IEEE Workshop on. — IEEE. 2012. — С. 25—32.

9. *Dubská M.*, *Herout A.*, *Sochor J.* Automatic Camera Calibration for Traffic Understanding. // BMVC. — 2014.

10. *Hoiem D.*, *Efros A. A.*, *Hebert M.* Putting objects in perspective // International Journal of Computer Vision. — 2008. — Т. 80, № 1. — С. 3—15.

11. *Viola P.*, *Jones M.* Rapid object detection using a boosted cascade of simple features // Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on. Т. 1. — IEEE. 2001. — С. I—511.

12. *Bourdev L.*, *Brandt J.* Robust object detection via soft cascade // 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Т. 2. — IEEE. 2005. — С. 236—243.

13. *Dollár P.*, *Belongie S.*, *Perona P.* The Fastest Pedestrian Detector in the West. // BMVC. Т. 2. — Citeseer. 2010. — С. 7.

14. *Dollár P.*, *Appel R.*, *Kienzle W.* Crosstalk cascades for frame-rate pedestrian detection // Computer Vision–ECCV 2012. — Springer, 2012. — С. 645—659.

15. *Krizhevsky A.*, *Sutskever I.*, *Hinton G. E.* Imagenet classification with deep convolutional neural networks // Advances in neural information processing systems. — 2012. — С. 1097—1105.

16. *Simonyan K.*, *Zisserman A.* Very deep convolutional networks for large-scale image recognition // arXiv preprint arXiv:1409.1556. — 2014.

17. Deep residual learning for image recognition / K. He [и др.] // arXiv preprint arXiv:1512.03385. — 2015.

18. Rethinking the inception architecture for computer vision / C. Szegedy [и др.] // arXiv preprint arXiv:1512.00567. — 2015.

19. Rich feature hierarchies for accurate object detection and semantic segmentation / R. Girshick [и др.] // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2014. — С. 580—587.

20. *Girshick R.* Fast r-cnn // Proceedings of the IEEE International Conference on Computer Vision. — 2015. — С. 1440—1448.

21. Faster R-CNN: Towards real-time object detection with region proposal networks / S. Ren [и др.] // Advances in neural information processing systems. — 2015. — С. 91—99.

22. Computer vision for interactive computer graphics / W. T. Freeman [и др.] // IEEE Computer Graphics and Applications. — 1998. — Т. 18, № 3. — С. 42—53.

23. *Isard M.*, *Blake A.* Condensation—conditional density propagation for visual tracking // International journal of computer vision. — 1998. — Т. 29, № 1. — С. 5—28.

24. Good features to track / J. Shi [и др.] // Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on. — IEEE. 1994. — С. 593—600.

25. *Kolsch M.*, *Turk M.* Fast 2d hand tracking with flocks of features and multi-cue integration // Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on. — IEEE. 2004. — С. 158—158.

26. *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.

27. (MP)2t: Multiple people multiple parts tracker / H. Izadinia [и др.] // European Conference on Computer Vision. — Springer. 2012. — С. 100—114.

28. *Yoon J. H.*, *Kim D. Y.*, *Yoon K.-J.* Visual tracking via adaptive tracker selection with multiple features // European Conference on Computer Vision. — Springer. 2012. — С. 28—41.

29. *Choi W.*, *Savarese S.* A unified framework for multi-target tracking and collective activity recognition // European Conference on Computer Vision. — Springer. 2012. — С. 215—230.

30. *Leal-Taixé L.*, *Pons-Moll G.*, *Rosenhahn B.* Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker // Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. — IEEE. 2011. — С. 120—127.

31. *Butt A. A.*, *Collins R. T.* Multi-target tracking by lagrangian relaxation to min-cost network flow // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2013. — С. 1846—1853.

32. *Andriyenko A.*, *Schindler K.*, *Roth S.* Discrete-continuous optimization for multi-target tracking // Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. — IEEE. 2012. — С. 1926—1933.

33. *Milan A.*, *Schindler K.*, *Roth S.* Detection-and trajectory-level exclusion in multiple object tracking // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2013. — С. 3682—3689.

34. Multi-hypothesis motion planning for visual object tracking / H. Gong [и др.] // 2011 International Conference on Computer Vision. — IEEE. 2011. — С. 619—626.

35. Coupling detection and data association for multiple object tracking / Z. Wu [и др.] // Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. — IEEE. 2012. — С. 1948—1955.

36. To track or to detect? an ensemble framework for optimal selection / X. Yan [и др.] // European Conference on Computer Vision. — Springer. 2012. — С. 594—607.

37. *Yang Y.*, *Ramanan D.* Articulated pose estimation with flexible mixtures-of-parts // Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. — IEEE. 2011. — С. 1385—1392.

38. *Felzenszwalb P.*, *McAllester D.*, *Ramanan D.* A discriminatively trained, multiscale, deformable part model // Computer Vision and Pattern

Recognition, 2008. CVPR 2008. IEEE Conference on. — IEEE. 2008. — С. 1—8.

39. *Pirsiavash H.*, *Ramanan D.* Steerable part models // Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. — IEEE. 2012. — С. 3226—3233.

40. Poselet conditioned pictorial structures / L. Pishchulin [и др.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2013. — С. 588—595.

41. Parsing occluded people / G. Ghiasi [и др.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2014. — С. 2401—2408.

42. *Chen X.*, *Yuille A. L.* Parsing occluded people by flexible compositions // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2015. — С. 3945—3954.

43. Modeling Instance Appearance for Recognition–Can We Do Better Than EM? / A. Chou [и др.] // International Workshop on Structured Prediction: Tractability, Learning, and Inference. — 2013.

44. *Finley T.*, *Joachims T.* Training structural SVMs when exact inference is intractable // Proceedings of the 25th international conference on Machine learning. — ACM. 2008. — С. 304—311.

45. Deformable part models are convolutional neural networks / R. Girshick [и др.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2015. — С. 437—446.

46. *Chen X.*, *Yuille A. L.* Articulated pose estimation by a graphical model with image dependent pairwise relations // Advances in Neural Information Processing Systems. — 2014. — С. 1736—1744.

47. Joint training of a convolutional network and a graphical model for human pose estimation / J. J. Tompson [и др.] // Advances in neural information processing systems. — 2014. — С. 1799—1807.

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

49. *Toshev A.*, *Szegedy C.* Deeppose: Human pose estimation via deep neural networks // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2014. — С. 1653—1660.

50. Efficient object localization using convolutional networks / J. Tompson [и др.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2015. — С. 648—656.

51. *Bulat A.*, *Tzimiropoulos G.* Human pose estimation via convolutional part heatmap regression // European Conference on Computer Vision. — Springer. 2016. — С. 717—732.

52. Building Statistical Shape Spaces for 3D Human Modeling / L. Pishchulin [и др.] // arXiv. — 2015. — Март.

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

54. *Kingma D.*, *Ba J.* Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. — 2014.

55. *Thirde D.*, *Li L.*, *Ferryman F.* Overview of the PETS2006 challenge // Proc. 9th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2006). — 2006. — С. 47—50.

56. *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.

57. Caffe: Convolutional Architecture for Fast Feature Embedding / Y. Jia [и др.] // arXiv preprint arXiv:1408.5093. — 2014.

58. *Шальнов Е. В.*, *Конушин А. С.* Использование геометрии сцены для увеличения точности детекторов // Программные продукты и системы. — 2017. — Т. 30, № 1. — С. 106—111.

59. *Tomasi C.*, *Kanade T.* Detection and tracking of point features. — 1991.

60. *Fulkerson B.*, *Vedaldi A.*, *Soatto S.* Class segmentation and object localization with superpixel neighborhoods // Computer Vision, 2009 IEEE 12th International Conference on. — IEEE. 2009. — С. 670—677.

61. *Bernardin K.*, *Stiefelhagen R.* Evaluating multiple object tracking performance: the CLEAR MOT metrics // EURASIP Journal on Image and Video Processing. — 2008. — Т. 2008, № 1. — С. 1—10.

62. *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.

63. *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.

64. *Ferrari V.*, *Marin-Jimenez M.*, *Zisserman A.* Progressive search space reduction for human pose estimation // Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. — IEEE. 2008. — С. 1—8.

65. *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.

Список рисунков

Список таблиц