НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

НАУЧНЫЙ ДОКЛАД

по результатам подготовленной
научно-квалификационной работы (диссертации)

«Нейросетевая модель определения позы и идентификации личности человека в видео»

ФИО Соколова Анна Ильинична

Направление подготовки 09.06.01 Информатика и вычислительная техника

Профиль (направленность) программы 05.13.18 – Математическое моделирование, численные методы и комплексы программ

Аспирантская школа по компьютерным наукам

Аспирант _____/Соколова А.И. /
подпись

Научный руководитель _____/Конушин А.С. /
подпись

Директор Аспирантской школы по компьютерным наукам

_____/Объедков С.А. /
подпись

Москва, 2019

1 Актуальность исследования

Согласно работам А. Маслоу, потребность в безопасности – одна из фундаментальных потребностей человека. Людям свойственно стремление обезопасить себя, защитить свое жилье от незаконного вторжения, а имущество – от хищений. С развитием современных систем видеонаблюдения появляется возможность фиксировать все происходящее на определенной территории и затем анализировать полученные данные. По видеозаписям можно отслеживать любые перемещения людей, определять незаконное проникновение на частную территорию, идентифицировать преступников, попадающих в поле зрения камер, контролировать доступ на закрытые объекты. Так, например, системы видеонаблюдения помогают при поимке взломщиков, грабителей или поджигателей, а также, будучи внедренными в распространненые сейчас системы домашней автоматизации («умный дом»), могут различать членов семьи и изменять поведение в зависимости от того, кто находится в кадре.

В последнее время особенно актуальной становится задача распознавания человека в видео. Личность человека определима в видео по нескольким признакам, наиболее точный из которых в настоящее время – черты лица. Однако качество распознавания сегодня позволяет доверить принятие решений машине только в кооперативном режиме – при сличении с фотографией в паспорте (обычно достаточно высокого качества). В реальной жизни (особенно при совершении преступлений) лицо человека может быть скрыто или плохо видно из-за недостаточного освещения, наличия головного убора, маски, грима и др. В этом случае распознавание возможно по другому признаку – походке. Согласно биометрическим исследованиям, походка индивидуальна и не поддается фальсификации, что делает ее уникальным идентификатором, сравнимым с отпечатками пальцев или радужной оболочкой глаз. Кроме того, в отличие от этих “классических” способов идентификации, походку можно наблюдать на большом расстоянии, без непосредственного контакта с человеком, поэтому в ряде случаев она служит единственным признаком для определения личности.

Задача распознавания человека по походке очень специфична в силу наличия множества факторов, меняющих походку визуально (каблуки или неудобная обувь; переносимые тяжелые предметы; одежда, скрывающая

части тела или сковывающая движения; скорость ходьбы) или влияющих на внутреннее представление модели походки (ракурс, освещение, различные параметры камеры). В связи с этим качество и надежность идентификации по походке гораздо ниже, чем по лицу, и, несмотря на успехи современных методов компьютерного зрения, данную задачу пока нельзя назвать решенной. Многие методы заточены исключительно под условия, присутствующие в базах данных, на которых они обучаются, что ограничивает их применимость в реальной жизни.

Помимо классических камер видеонаблюдения, сохраняющих все происходящее в кадре $25-30$ раз в секунду, в последние годы набирают популярность другие виды сенсоров, в частности сенсоры динамического зрения (Dynamic Vision Sensors, DVS).

В отличие от видеокамер, сенсор, подобно сетчатке глаза, фиксирует изменения интенсивности в каждом пикселе, игнорируя точки с постоянной яркостью. В условиях статического сенсора события в точках фона генерируются очень редко, предотвращая хранение излишней информации. Одновременно с этим интенсивность в каждой точке измеряется несколько тысяч раз в секунду, что приводит к асинхронному захвату всех важных изменений. В результате такой поток событий оказывается очень информативным и годится для получения данных, необходимых для решения множества задач видеоанализа, требующих извлечения динамических характеристик, в том числе распознавания по походке.

Еще одно преимущество потока событий перед классическим видеорядом – его чувствительность к изменению интенсивности пикселей. Точки цифровых изображений кодируются числами от 0 до 255 и имеют дискретный спектр значений, а сенсор динамического зрения генерирует события при мельчайших изменениях яркости. Это позволяет применять его даже в условиях плохого освещения, когда движение объектов едва различимо. Упомянутые преимущества делают DVS сенсоры перспективной быстро развивающейся технологией, что приводит к необходимости решения задач видеоанализа для получаемых данных. Несмотря на постоянное развитие методов компьютерного зрения, до сих пор не было предложено подходов к решению задачи распознавания по походке по данным сенсоров динамического зрения, и она представляет собой обширное поле для исследований.

С развитием методов компьютерного зрения появляется множество подходов к идентификации человека по движениям в видео. Эти подходы состоят в использовании различных естественных биометрических характеристик (поза человека, его силуэт, их изменение во время ходьбы). Однако поскольку распознавание предполагается именно по движениям, а не по форме, фигуре или другим внешним характеристикам, в качестве основного источника информации предлагается рассматривать оптический поток между соседними кадрами видео. Благодаря тому что оптический поток содержит информацию исключительно о сдвигах точек и не зависит от яркости или контрастности кадров, влияние внешних характеристик при таком подходе уменьшается, что позволяет развить метод идентификации человека по движениям, дополняющий другие подходы к распознаванию (распознавание по лицу или реидентификацию).

Наиболее успешными в решении задач компьютерного зрения в последние годы стали методы глубинного обучения, основанные на обучении нейронных сетей. Признаки, обучаемые с помощью нейронных сетей, часто обладают более высоким уровнем абстракции, необходимым для качественного распознавания. Это позволило добиться выдающихся результатов в решении таких задач, как классификация видео и изображений, сегментация изображений, детекция объектов, визуальный трекинг и др. Однако, несмотря на успешность методов глубинного обучения, в задаче распознавания походки неглубокие модели все еще опережают нейросетевые на некоторых наборах данных, и оба подхода не достигают приемлемой точности.

2 Цель и задачи исследования

Целью диссертационного исследования является разработка и программная реализация нейросетевого алгоритма идентификации человека в видео на основании движения точек фигуры человека, устойчивого к смене ракурса съемки и наличию переносимых предметов.

Для достижения этой цели были поставлены следующие задачи:

1. Разработать и реализовать алгоритм распознавания человека в видео, анализирующий оптический поток;

2. Разработать и реализовать метод многоракурсного распознавания человека по походке, основанный на анализе движения разных частей тела человека;

3. Разработать алгоритм распознавания человека по походке для данных, получаемых с сенсора динамического зрения.

3 Формальная постановка задачи

Формальным объектом исследования являются видеопоследовательности v с камер видеонаблюдения и потоки событий e с сенсоров динамического зрения с движущимся в них человеком. При наличии размеченной галереи D видеофрагментов движущихся людей требуется определить, кем из галереи является человек в видео, то есть предсказать личность x человека.

Пусть галерея задана в виде

$$D = \{(v_i, x_i)\}_{i=1}^{N},\ x_i \in P$$

где N – общее количество записей, P – множество людей. Для исследуемого видеофрагмента v требуется найти метку объекта $x \in P$. Задача сводится к нахождению некоторой метрики близости S, по которой ищется ближайший к данному объект в галерее.

$$S(v, v_i) \to \min_{v_i:\ \exists x_i:\ (v_i, x_i) \in D}$$

Для потоков событий задача формулируется аналогично.
При этом на все видеофрагменты накладывается набор ограничений:

- в каждом видео в галерее присутствует ровно один человек;
- человек виден в полный рост;
- отсутствуют перекрытия;
- камера статична;
- набор возможных ракурсов (высота камеры и ее наклон) ограничен.

Данные условия вводятся в связи с ограниченностью существующих эталонных наборов данных.

4 Степень разработанности темы исследования

Задача анализа движений человека вызывала интерес исследователей во многих областях (психологии, медицине, биометрии) задолго до развития методов компьютерного зрения. Первые работы по распознаванию походки по последовательности изображений появились в 1970-е годы, однако именно с развитием методов машинного зрения задача идентификации по походке получила активное развитие и стала актуальной во множестве областей. Первым из наиболее успешных подходов стал метод, основанный на изображениях энергии походки (Gait Energy Images, GEI [7]), предложенный в 2006 году. Такие изображения – усредненные по одному циклу походки бинарные маски силуэта движущегося человека. Они характеризуют частоты нахождения человека в той или иной позе во время движения. Несмотря на простоту, этот метод получил широкое распространение, и до сих пор многие современные подходы базируются именно на изображениях энергии. Производя агрегацию сырых данных, вычисление GEI карт позволяет получить один дескриптор для всего видеоряда, к которому впоследствии могут быть применены любые методы машинного зрения. Наиболее успешные на сегодняшний день подходы рассматривают в качестве исходных данных именно изображения энергии походки. Так, Ли [11] предлагает байесовский вероятностный подход, рассматривая GEI изображения как случайные матрицы и считая вариативность ракурса и внешних условий шумом, прибавляемым к сущности походки.

Другой популярный источник информации о движениях человека – траектории точек. Кастро в своей работе [1] рассматривает траектории точек фигуры и по ним строит дескрипторы Фишера, характеризующие походку человека. Также для распознавания по движениям часто используется информация о позе человека в каждом кадре и ее изменении. Поза как правило характеризуется положением скелета человека – набором положений основных суставов и ключевых точек фигуры, и именно положения и относительные сдвиги и скорости ключевых точек предлагают использовать для распознавания в [5, 19].

Работа [11] предложена в 2017 году, уже после появления и начала активного развития современных методов глубинного обучения, однако полученное в ней качество распознавания не уступает нейросетевым методам.

Распознавание по походке в видео до сих пор остается одной из немногих задач, в которых появляются новые неглубокие алгоритмы, достигающие качества глубоких.

Однако стоит отметить, что нейронные сети, достигающие высочайших успехов в большинстве задач компьютерного зрения, таких как классификация изображений, сегментация, детекция, оценка глубины, распознавание действий и т.д., также широко применяются для идентификации человека по движениям. Глубинные методы распознавания походки можно разделить на два типа: классификационные [2, 13] и верификационные [17, 20]. Первые обучаются предсказывать вероятность принадлежности видео тому или иному субъекту, вторые оценивают, присутствует в двух видео один и тот же человек или разные. Второй подход оказывается более общим, так как для перехода от верификации к классификации не требуется никаких дополнительных алгоритмов и оценок, необходимых для перехода в обратную сторону. До сих пор очень многие нейросетевые модели используют в качестве входных данных отдельные силуэты [3, 25] или уже упомянутые изображения энергии походки [13]. Основные различия заключаются в выборе архитектур (однопоточных для классификации [13] и сиамских двух- и трехпоточных для верификации [17, 20, 24]), методов сравнения дескрипторов, их агрегации. Особое внимание к агрегации признаков уделяется в [3], где набор признаков, полученных из силуэтов одного человека при разных условиях, рассматривается как неупорядоченное множество, к которому применяется комбинация различных статистических функций. Кроме того, к задаче распознавания по походке оказываются применимы и современные генеративные подходы [8, 21, 22], используемые для преобразования изображения энергии, полученного под одним углом, к другому, более удобному для распознавания.

Другими входными данными для нейросетевых моделей могут выступать, оптический поток [2], поза человека, характеризующаяся положением его основных суставов [6], и др. Однако, как уже было сказано, наиболее популярной формой данных остаются силуэты, и именно их используют в самых успешных на сегодняшний день моделях идентификации по походке.

В настоящее время существует несколько общедоступных наборов данных для распознавания человека по походке, используемых для обучения и сравнения методов. Наиболее популярные из них – коллекции TUM-

GAID [9], CASIA Gait dataset B [23] и OU-ISIR gait database [10, 18]. База TUM создана для распознавания сбоку: она содержит видео двигающихся вдоль стены людей, снятых под углом 90°. Для каждого из 305 человек в коллекции присутствует 10 видео, среди которых 2 – с рюкзаком и 2 – в сменной обуви. Набор CASIA гораздо меньше по количеству людей: он содержит данные для 124 человек, однако каждый человек снят с 11 разных точек под углами от 0° до 180° (с шагом 18°), что делает этот набор сложным для распознавания. Еще два набора данных – коллекции университета Осаки OULP [10] и OUMVLP [18] – самые большие из существующих наборов. Они содержат порядка 4 тысяч и 10 тысяч людей соответственно. Данные OULP сняты под углом от 55° до 85°, OUMVLP – от 0° до 90° и от 180° до 270°. Однако названные коллекции распространяются исключительно в форме бинарных масок силуэтов, поэтому не все существующие методы могут быть обучены или протестированы на них. Тем не менее те алгоритмы, которые используют силуэты в качестве источника данных, сравниваются именно на этих коллекциях.

5 Научная новизна

1. Впервые предложен и реализован метод распознавания человека по походке, основанный на исследовании движения точек в различных областях фигуры человека.

2. Предложен оригинальный метод повышения устойчивости алгоритма к смене ракурса путем регуляризации модели и проецирования в специальное признаковое пространство, снижающее зависимость от ракурса.

3. Выполнено оригинальное исследование переносимости алгоритмов распознавания по походке между различными наборами данных.

4. Впервые предложен метод распознавания человека по движениям в данных, получаемых с сенсора динамического зрения.

6 Практическая значимость

Предложенный алгоритм распознавания человека в видео может быть внедрен в системы домашней автоматизации («умный дом»), распознающие членов семьи и изменяющие свое поведение в зависимости от того, кто присутствует в кадре. Будучи объединенной с сигнализацией, система может реагировать на появление людей, не входящих в состав семьи, и отслеживать незаконное проникновение в частные дома.

Кроме того, алгоритм идентификации по походке может быть использован в местах крупных скоплений людей, таких как вокзалы и аэропорты, где нет возможности производить съемку крупным планом, но есть очевидная необходимость отслеживания и контроля доступа.

7 Основные результаты исследования и положения, выносимые на защиту

1. Предложен и реализован метод распознавания человека по походке сбоку, анализирующий последовательные сдвиги точек между соседними кадрами видео.

2. Предложен и реализован метод многоракурсного распознавания человека по походке, основанный на рассмотрении движения точек в различных областях фигуры человека.

3. Выявлено влияние движения точек в различных частях фигуры на качество распознавания.

4. Предложен и реализован алгоритм распознавания по походке, показывающий устойчивость к переносу между различными коллекциями данных.

5. Предложен и реализован оригинальный метод повышения устойчивости алгоритма к смене ракурса.

6. Предложен и реализован метод распознавания человека по движениям в данных, получаемых с сенсора динамического зрения.

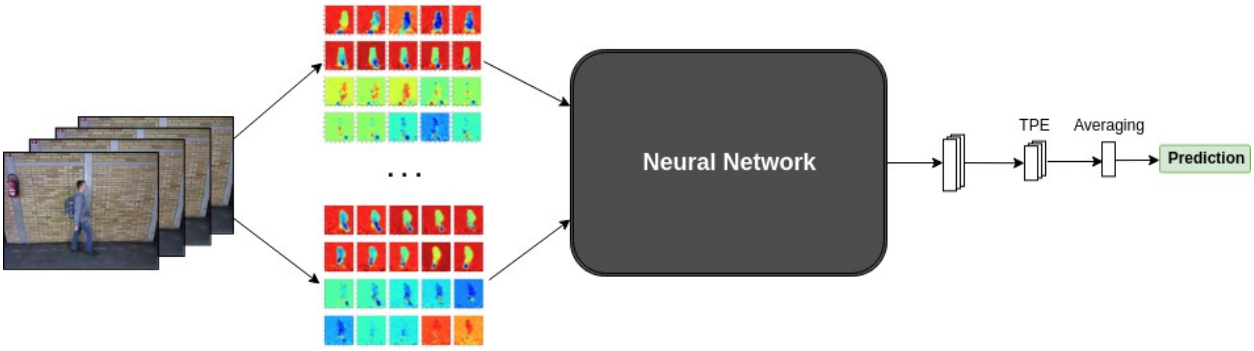


Рис. 1: Схема базового алгоритма распознавания.

8 Содержание работы

Во **введении** обоснована актуальность диссертационных исследований, поставлены цель, задачи, а также сформулированы научная новизна и практическая значимость работы.

Первая глава посвящена обзору существующих методов решения задачи распознавания по походке: описаны их достоинства и недостатки, обосновывающие актуальность данной диссертационной работы.

Во **второй главе** предлагается базовый алгоритм решения задачи идентификации человека по походке в видео с бокового ракурса. Глава начинается с неформальной постановки задачи и выявления основных сложностей, препятствующих построению успешного алгоритма.

Далее следует описания формального алгоритма решения задачи и разделы, соответствующие каждому из его этапов. Предложенный алгоритм состоит из

1. Предобработки данных,
2. Извлечения признаков походки путем вычисления скрытых представлений нейронной сети,
3. Постобработки извлеченных признаков и их агрегации в один дескриптор походки для всего видео,
4. Классификации дескрипторов походки.

Схема базового подхода изображена на рис. 1.

В разделе 2.1 описывается предложенный алгоритм предобработки данных.

В качестве основного источника информации, из которого впоследствии извлекаются признаки, предлагается использовать оптический поток, т.е. карты сдвигов точек между парами последовательных кадров. Оптический поток в совокупности с признаками, полученными из цветных кадров напрямую, показал высокие результаты в задаче распознавания действий [14]. Это дает повод полагать, что полученных данных действительно достаточно для качественного анализа видео данных.

Чтобы учтывать не только "мгновенные" сдвиги точек между парами соседних кадров, но и продолжительное движение, предлагается рассматривать не отдельные карты оптического потока, а блоки, составленные из нескольких идущих подряд карт. Такие блоки совмещают в себе краткосрочные и долгосрочные динамические характеристики походки.

Полный алгоритм предобработки выглядит следующим образом:

1. В каждом кадре видео детектируется человек.
2. Между парами соседних кадров вычисляется оптический поток.
3. Карты объединяются в блоки по 10 кадров с пересечением по 5 кадров.
4. Для каждого блока вычисляется рамка, содержащая фигуру человека в каждом кадре. Из блоков вырезается прямоугольник в соответствии с этой рамкой.

Раздел 2.2 посвящен исследованию влияния архитектуры нейронной сети на качество распознавания. В качестве базовой модели была выбрана классическая архитектура CNN [4], первой показавшая высокие результаты решения многих задач компьютерного зрения. Однако за годы развития методов глубинного обучения было предложено множество различных подходов к построению нейронных сетей. Помимо базовой модели в данном разделе исследуется слияние нескольких потоков сиамской сети и модель из семейства VGG[15]).

Нейронная сеть во всех случаях обучается методом стохастического градиентного спуска для задачи классификации, получая на вход тензоры размера $(2N \times W \times H)$ и возвращая вектор распределения вероятностей принадлежности к классам (размера, равного количеству людей в обучающей

выборке). Сеть обучается один раз на фиксированном наборе данных и впоследствии используется для извлечения признаков походки – выходов последнего скрытого слоя низкой размерности.

Такой подход позволяет использовать одну и ту же модель без дообучения при появлении в базе новых людей. Меняться в этом случае будет только модель, предсказывающая класс по выделенным признакам, дообучение или перенастройка которого, как правило, требует гораздо меньших временных затрат, чем обучение нейронной сети.

В разделе 2.3 описывается метод обработки признаков, извлеченных из обученной нейронной сети. Несмотря на то, что признаки уже имеют невысокую размерность, дополнительное ее уменьшение путем применения метода главных компонент (PCA) приводит к избавлению от оставшегося в признаках шума и повышению качества распознавания. Кроме того, уменьшение размерности ускоряет последующую классификацию, делая весь алгоритм более эффективным.

Однако PCA преобразование – не единственная предложенная модификация признаков. Развивая идею уменьшения размерности, я предлагаю обучить вложение [12] признаков в пространство более низкой размерности. Это вложение обучается таким образом, чтобы степень близости признаков одного объекта оказалась выше, чем разных. Требуемое вложение задается матрицей проекции W на искомое пространство, для нахождения которой решается следующая задача оптимизации. Определим функцию S_W близости векторов, т.ч. для любой тройки (v_a, v_p, v_n) где v_a и v_p принадлежат одному классу, а v_n – другому, будет выполнено

$$S_W(v_a, v_p) > S_W(v_a, v_n) \tag{1}$$

Функция S_W параметризуется матрицей W: $S_W(v, u) = (Wv)^T(Wu)$. Тогда задача нахождения S_W сводится к оптимизации триплетной функции ошибки, решаемой методом градиентного спуска.

Для своей задачи я трансформировала этот метод следующим образом: вместо косинусной меры близости (и рассмотрения скалярного произведения в качестве S_W) используется Евклидово расстояние, а знак неравенства и 1 меняется на противоположный. Решение оптимизационной задачи, как и в изначальной формулировке, происходит методом стохастического градиентного спуска с подбором сложных обучающих примеров.

Применяя описанное вложение и усредняя полученные дескрипторы по всем блокам, я получаю финальные дескрипторы походки, которые можно классифицировать, предлагаемый метод описывается в разделе 2.4. Исследуются различные метрики близости дескрипторов и выявляется наиболее подходящая для задачи распознавания по походке.

Раздел 2.5 посвящен описанию используемых коллекций походок. Основными наборами данных, послужившими для экспериментальной оценки предложенного метода, являются базы TUM GAID [9] и CASIA Gait Dataset B [23]. Первый из них использовался для исследования распознавания сбоку, а второй, обладая большой вариативностью углов съемки, – для многоракурсной идентификации. Кроме того, наличие нескольких наборов данных позволило провести эксперименты по переносимости алгоритмов между коллекциями.

Экспериментальная оценка и подробный протокол тестирования предложенного метода приведены в разделе 2.6. В таблице 1 сравнивается качество моделей, использующих разные архитектуры, а также способы обработки нейросетевых дескрипторов. Самых высоких результатов удалось достичь при использовании VGG архитектуры, наиболее успешной на момент исследований.

Модель	Качество распознавания	
Архитектура, пост-обработка, мера близости	Rank-1	Rank-5
CNN (PCA 1100), L_1	93,22	98,06
CNN (PCA 1100), L_2	92,79	98,06
CNN+TPE (PCA 450), L_2	94,51	98,70
fusion CNN (PCA 160), L_1	93,97	98,06
fusion CNN (PCA 160), L_2	94,40	98,06
fusion CNN +TPE (PCA 160), L_1	94,07	98,27
fusion CNN +TPE (PCA 160), L_2	95,04	98,06
VGG (PCA 1024), L_1	97,20	99,78
VGG (PCA 1024), L_2	96,34	99,67
VGG+TPE (PCA 800), L_1	**97,52**	**99,89**
VGG+TPE (PCA 800), L_2	96,55	99,78

Таблица 1: Сравнение архитектур и методов классификации на наборе TUM-GAID

Кроме того, в этом разделе приведены эксперименты по переносу алгоритма между базами. В связи с тем что вариативность походки очень велика и единой базы данных, учитывающей все возможные условия, не

существует, модели переобучаются под условия, представленные в обучающей коллекции. Мной впервые был проведен эксперимент по переносу модели между коллекциями и совместному обучению на нескольких наборах для повышения качества распознавания.

В таблице 2 приведено качество переноса наилучшей модели между базами. При неизменном подходе к обучению изменяются обучающая и тестовая выборки и исследуется переобучение модели под набор данных.

Тестовая выборка / обучающая выборка	CASIA	TUM
CASIA	74,93%	67,41%
TUM	58,20%	97,20%
Объединение коллекций CASIA + TUM	72,06%	96,45%

Таблица 2: Результаты переноса моделей между наборами данных

Как видно из таблицы, наличие объектов из базы в обучающей выборке существенно повышает качество идентификации. Это говорит о том, что коллекции, будучи внешне схожими, отличаются и модели переобучаются под свойства каждой из них.

Третья глава посвящена следующему этапу исследования – распознаванию на основе движения точек в различных частях тела человека. Общая схема такого метода распознавания изображена на рис. 3. Глава начинается с обоснования необходимости рассмотрения отдельных частей тела человека, а не только фигуры в целом.

В раделе 3.1 описываются исследуемые области фигуры человека и обосновывается, почему выбор падает именно на них. В наборе рассматриваемых областей присутствуют части фигуры разного размера: от фигуры целиком до небольших областей вокруг суставов, что дает возможность получать признаки разного масштаба и уделять некоторым областям повышенное внимание.

Раздел 3.2 посвящен основному “телу” алгоритма – обучению нейронной сети. Входными данными сети также являются карты оптического потока, однако из каждой карты берется несколько областей, соответствующих выбранным частям тела (рис. 2). Для всех выбранных областей обучается одна и та же сверточная нейронная сеть, предсказывающая распределение вероятностей классов по отдельным картам оптического потока. Как

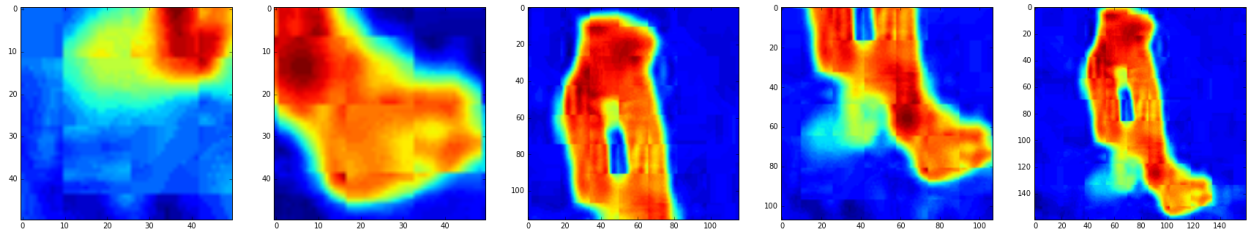


Рис. 2: Визуализация оптического потока вблизи различных частей тела человека (слева направо): левая стопа, правая стопа, верхняя половина, нижняя половина и фигура целиком.

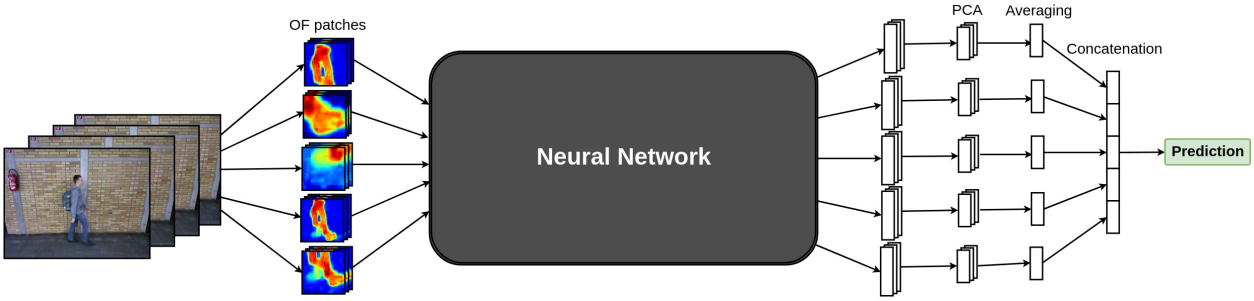


Рис. 3: Схема алгоритма распознавания на основе движения частей тела человека.

и в главе 2, исследуется влияние архитектуры на качество распознавания, однако базовой моделью выбрана VGG, показавшая лучший результат в одноракурсном методе. Лучшей, однако, оказывается WideResNet архитектура, основанная на так называемых остаточных соединениях (residual connections).

В разделе 3.3 исследуются методы постобработки и агрегации нейросетевых признаков. Поскольку из каждого кадра вырезается несколько областей, оказывается столько же извлеченных дескрипторов. Наивная агрегация путем усреднения признаков и по времени, и по частям тела дает приемлемый результат, однако более успешным оказывается временное усреднение, когда вычисляется среднее представление по всему видео для каждой части тела, и полученные дескрипторы конкатенируются в один высоко-размерный вектор. Такой способ не смешивает все признаки и позволяет сохранить информацию, извлеченную с разных масштабов и из разных областей тела.

Раздел 3.4 посвящен экспериментальной оценке предложенного подхода. Сравниваются модели, использующие различные наборы частей тела, ар-

хитектуры и методы агрегации. Наилучшего качества удается добиться при помощи WideResNet архитектуры и временного усреднения. Зависимость точности идентификации от набора частей тела приведена в таблице 3.

Таблица 3: Сравнение качества распознавания на базе TUM-GAID моделей, использующих различные части тела.

Части тела	Точность
Стопы	79,7%
Нижняя половина тела	96,3%
Верхняя половина тела	96,2%
Фигура целиком	98,9%
Фигура целиком, верхняя и нижняя половины тела	99,4%
Фигура целиком, верхняя и нижняя половины тела, стопы	99,8%

Как показывают эксперименты (таблица 3), добавление в модель более мелких частей тела действительно повышает качество идентификации.

Кроме того, в данном разделе приведена экспериментальная оценка многоракурсного распознавания. Преимущество предложенного подхода состоит в независимости от ракурса съемки: он может быть применим к данным, снятым под любым углом. В таблице 4 отражена усредненная точность кросс-ракурсного распознавания и проведено ее сравнение с лучшими из существующих методов.

Таблица 4: Сравнение средней точности распознавания для трех тестовых углов базы CASIA

Метод	Средняя точность [%]			
Тестовый ракурс / Модель	54	90	126	Среднее
WideResNet (PCA 230), concat, L_1	**77,8**	**68,8**	74,7	**73,8**
SPAE [22]	63,3	62,1	66,3	63,9
Wu [20]	**77,8**	64,9	**76,1**	72,9

Четвертая глава посвящена методам повышения устойчивости алгоритмов к смене ракурсов. Вариативность углов съемки – наиболее сложная проблема распознавания по походке: ходьба одного и того же человека может выглядеть по-разному при съемке с различных ракурсов.

В разделе 4.1 предложена модификация процесса обучения модели для получения признаков походки, не зависящих от ракурса. Предлагается, не

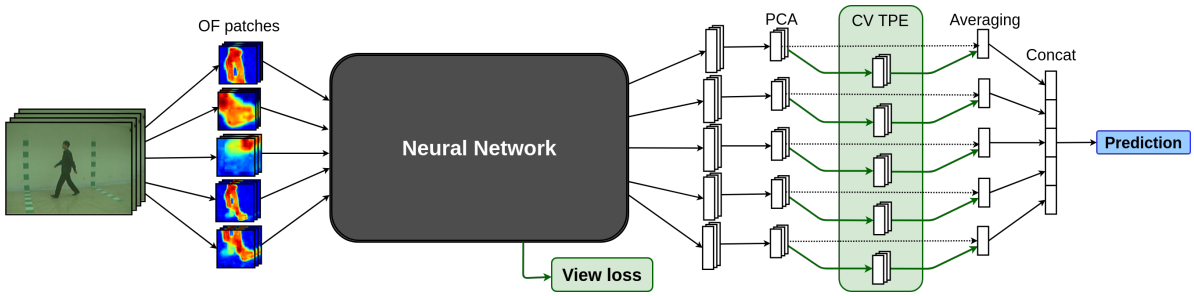


Рис. 4: Ракурсо-устойчивый алгоритм распознавания. Зеленым обозначены добавленные модули повышения стабильности: «ракурсная» функция потерь и многоракурсное триплетное вероятностное вложение CV TPE.

меняя архитектуру сети, добавить в нее дополнительную функцию ошибки, которая штрафует за то, что признаки походки одного и того же человека при фиксированном ракурсе отличаются от среднего признака походки по всем ракурсам. Такая «ракурсная» функция потерь, добавленная к кросс-энтропии, отвечающей за правильную классификацию, выступает в качестве регуляризатора и препятствует попаданию параметров в локальный минимум. В разделе приводится полный алгоритм обучения, чередующий шаги с регуляризацией и без нее, и показывается, что использование новой функции штрафа позволяет заметно уменьшить нерегуляризованные потери.

Раздел 4.2 посвящен еще одному методу преодоления зависимости от ракурса. Для признаков, извлеченных из нейронной сети, предлагается обучить многоракурсное триплетное вероятностное вложение (cross-view triplet probabilistic embedding, CV TPE), сближающее признаки одного объекта, полученные под разными углами, и отдаляющее друг от друга признаки разных объектов с совпадающим ракурсом. Такое вложение – модификация метода, используемого в базовом подходе и описанного в разделе 2.3, однако добавление зависимости от ракурса при сэмплировании троек объектов значительно влияет на качество многоракурсного распознавания и улучшает его сильнее классического вложения. Схема ракурсо-устойчивого алгоритма приведена на Рис. 4.

Результаты экспериментов и сравнение с наилучшими на момент исследования методами приведено в разделе 4.3. Можно сделать вывод (см. таблицу 5), что совместное использование двух предложенных подходов оказывает больший эффект на качество распознавания, чем их использование

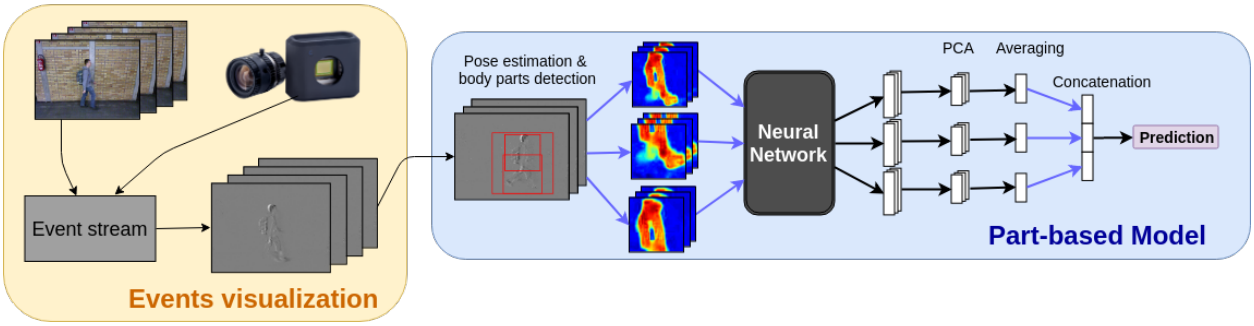


Рис. 5: Алгоритм распознавания данных с сенсора динамического зрения

по-отдельности. Это свидетельствует о том, что рассмотренные подходы дополняют друг друга и действительно повышают устойчивость модели к смене ракурса.

Таблица 5: Точность ракурсо-устойчивого распознавания на наборе CASIA и сравнение с наилучшими методами.

Метод	Средняя точность [%]			
Тестовый ракурс / Модель	54	90	126	Среднее
Базовая модель на основе частей тела	77.8	68.8	74.7	73.77
Базовая модель + «ракурсный» штраф	83.6	74.6	83.7	80.63
Базовая модель + TPE	82.7	73.5	82.4	79.53
Базовая модель + CV TPE	84.6	75.6	84.4	81.53
Базовая модель + «ракурсный» штраф + CV TPE	86.3	**75.8**	**86.0**	**82.70**
SPAE [22]	63.3	62.1	66.3	63.90
Wu [20]	77.8	64.9	76.1	72.93
GaitSet [3]	**86.5**	75.5	**86.0**	82.67

Пятая глава посвящена исследованию распознаваемости человека в потоке событий и применимости к нему разработанных методов идентификации по походке. В первом разделе главы приводится обоснование необходимости решения задачи идентификации для рассматриваемых данных. Благодаря удобству и эффективности сенсоров динамического зрения они набирают все большую популярность и все чаще заменяют классические камеры в системах видеонаблюдения, что приводит к необходимости развития методов компьютерного зрения для подобных данных.

В разделе 5.2 объясняется необходимость визуализации потока событий и описываются способы, которыми это можно сделать. В силу того что

идентификация по походке – задача анализа движения человека, для ее решения важно иметь информацию о взаимном расположении точек фигуры и их относительных перемещениях. Кроме того, чтобы иметь возможность применять нейросетевые методы анализа видео, необходимо трансформировать набор событий в отдельных точках в двух- или трехмерные (в зависимости от количества каналов) тензоры, содержащие в себе все, что снято сенсором.

Простейший способ визуализации, позволяющий, однако, достичь приемлемого качества распознавания, заключается в суммировании событий в каждой точке в некотором временном интервале. При этом сумма вычисляется с учетом полярности события (1 или -1 в зависимости от того, увеличилась или уменьшилась яркость в точке). Кадры, полученные таким образом, легко интерпретируемы: точки, в которых в течение всего временного интервала ничего не происходит, имеют нулевое значение, а те, в которых появляются какие-либо изменения, окрашиваются в тот или иной цвет в зависимости от количества событий, произошедших в них за текущий промежуток времени, и их полярности. Таким образом, изображения можно рассматривать как меру событий в каждой точке. Примеры визуализаций потока событий в оттенках серого изображены на рис. 6, фигура действительно оказывается различима невооруженным глазом.

Раздел 5.3 посвящен изменениям, которые необходимо произвести с предложенным методом, чтобы применить его к визуализированным потокам событий. Важным отличием является отказ от рассмотрения стоп, т.к. в то время, когда человек делает шаг, его опорная стопа оказывается неподвижна и в течение некоторого интервала времени события в соответствующей области не генерируются, делая стопу невидимой.

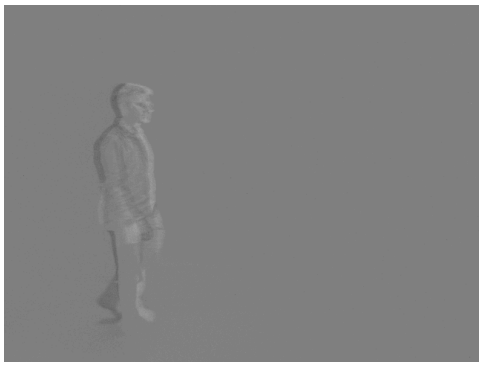
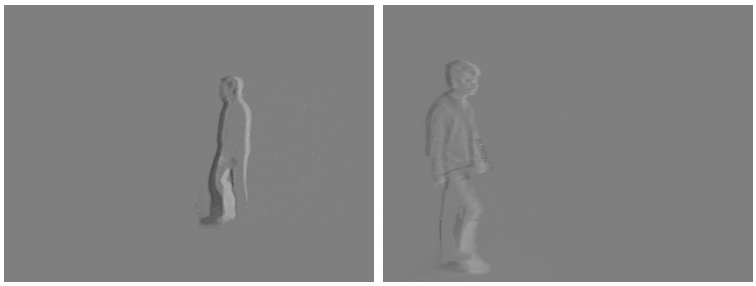


Рис. 6: Пример визуализации потока событий.

Кроме того, изменения претерпевают этапы детекции человека и оценки позы. Первая задача упрощается, т.к. в условиях статической камеры и отсутствия перекрытий фигура человека оказывается единственным объектом в некоторой окрестности, точки которого генерируют события (рис. 6), поэтому задача детекции сводится к задаче отделения фигуры человека от

монотонного фона, решаемой стандартными детекторами на основе изменения яркости пикселей. Задача оценки скелета человека, наоборот, оказывается более сложной: из-за визуальных отличий от обычных изображений существующие методы детекции ключевых точек фигуры не справляются с новыми данными без дополнительной настройки и требуют дообучения.

Раздел 5.4 посвящен данным, используемым в экспериментах. Ранее никто не исследовал возможность распознавания человека по походке в данных, полученных с DVS сенсора, поэтому выборок, напрямую подходящих для решения данной задачи, нет в открытом доступе. Одной из поставленных задач стал сбор собственной DVS-коллекции походок, однако для экспериментов был сгенерирован набор потоков событий из существующих баз походок. Генерация данных происходит в два этапа: интерполяция промежуточных кадров и сама генерация событий. В простейшем случае применяется попиксельная линейная интерполяция, которая, несмотря на простоту, дает качественный результат, визуально близкий к реальным данным. Последующая генерация событий происходит путем сравнения изменения яркости точек в соседних кадрах с заданным порогом. На рис. 7 изображены визуализации реальных и симулированных данных, показывающие их сходство.



а) б)

Рис. 7: а) визуализация симулированных данных, б) визуализация реальных данных.

В разделе 5.5 приводится экспериментальная оценка предложенного метода. Сравниваются модели, использующие разные части фигуры человека и получено численное доказательство (таблица 6) того, что области стоп человека в данной модели оказываются неинформативными и лишь добавляют шум, уменьшая точность идентификации.

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

1. Предложен и реализован метод распознавания человека по походке сбоку, анализирующий последовательные сдвиги точек между сосед-

Таблица 6: Results of the end-to-end model on simulated TUM-GAID dataset.

Набор частей тела	Rank-1 [%]	Rank-5 [%]
Фигура целиком	98,0	99,7
Фигура целиком, верхняя и нижняя половины тела	99,0	100,0
Фигура целиком, верхняя и нижняя половины тела, стопы	98,8	99,8
Original RGB model [16]	99,8	100

ними кадрами видео.

2. Предложен и реализован метод многоракурсного распознавания человека по походке, основанный на рассмотрении движения точек в различных областях фигуры человека.

3. Выявлено влияние движения точек в различных частях фигуры на качество распознавания.

4. Предложен и реализован алгоритм распознавания по походке, показывающий устойчивость к переносу между различными коллекциями данных.

5. Предложен и реализован оригинальный метод повышения устойчивости алгоритма к смене ракурса.

6. Предложен и реализован метод распознавания человека по движениям в данных, получаемых с сенсора динамического зрения.

Дальнейшее развитие исследуемой темы возможно по следующим направлениям:

1. Разработка и реализация метода оценки ракурса съемки походки;

2. Добавление информации об угле съемки в модель распознавания для улучшения качества идентификации и повышения устойчивости к смене ракурса;

3. Разработка и реализация метода синтеза многоракурсных RGB и DVS путем применения методов захвата движений и генерации трехмерных данных для расширения существующих обучающих и тестовых выборок;

4. Применения синтетических данных для повышения качества многоракурсного распознавания человека в различных типах видеоданных (классические RGB, DVS).

9 Апробация результатов исследования (конференции, научные публикации)

По результатам работы сделаны следующие доклады на конференциях

1. ISPRS INTERNATIONAL WORKSHOP "PHOTOGRAMMETRIC AND COMPUTER VISION TECHNIQUES FOR VIDEO SURVEILLANCE, BIOMETRICS AND BIOMEDICINE" – PSBB17, MOSCOW, RUSSIA, MAY 15-17, 2017 "Gait recognition based on convolutional neural networks"
2. 28-я Международная конференция по компьютерной графике и машинному зрению "ГрафиКон 2018", Томск, 24–27 сентября 2018 года "Обзор методов распознавания человека по походке в видео"
3. 16th International Conference on Machine Vision Applications (MVA), Tokyo, Japan, May 27-31, 2019 "Human identification by gait from event-based camera"

Основные результаты работы были опубликованы в следующих статьях:

1. A. Sokolova, A. Konushin, Gait recognition based on convolutional neural networks, In International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 2017 (Scopus)
2. A. Sokolova, A. Konushin, Pose-based deep gait recognition, In IET Biometrics, 2019, (Scopus, Q2)
3. A. Sokolova, A. Konushin, Human identification by gait from event-based camera, In MVA Proceedings, 2019 (Scopus)
4. A. Sokolova, A. Konushin, Methods of gait recognition in video, Programming and Computer Software, 2019 (WoS, Scopus, Q3)
5. А. Соколова, А. Конушин, Методы идентификации человека по походке в видео, Труды Института системного программирования РАН, том 31, № 1, с. 69-82

Список литературы

[1] F. Castro, M. Marín-Jiménez, and R. Medina-Carnicer. Pyramidal Fisher Motion for multiview gait recognition. In *22nd International Conference on Pattern Recognition*, pages 1692–1697, 2014.

[2] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. Pérez de la Blanca. Automatic learning of gait signatures for people identification. In I. Rojas, G. Joya, and A. Catala, editors, *Advances in Computational Intelligence*, pages 257–270, Cham, 2017. Springer International Publishing.

[3] H. Chao, Y. He, J. Zhang, and J. Feng. GaitSet: Regarding gait as a set for cross-view gait recognition. In *AAAI*, 2019.

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. BMVC*, 2014.

[5] M. Deng, C. Wang, F. Cheng, and W. Zeng. Fusion of spatial-temporal and kinematic features for gait recognition with deterministic learning. *Pattern Recognition*, 67:186 – 200, 2017.

[6] Y. Feng, Y. Li, and J. Luo. Learning effective gait features using LSTM. In *International Conference on Pattern Recognition*, page 325–330, 2016.

[7] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(2):316–322, 2006.

[8] Y. He, J. Zhang, H. Shan, and L. Wang. Multi-task gans for view-specific feature learning in gait recognition. *IEEE Transactions on Information Forensics and Security*, 14(1):102–113, 2019.

[9] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll. The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits. *J. of Visual Com. and Image Repres.*, 25(1):195 – 206, 2014.

[10] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi. The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait Recognition. *IEEE Trans. on Information Forensics and Security*, 7, Issue 5:1511–1521, Oct. 2012.

[11] C. Li, S. Sun, X. Chen, and X. Min. Cross-view gait recognition using joint Bayesian. In *Proc. SPIE 10420, Ninth International Conference on Digital Image Processing (ICDIP 2017)*, 2017.

[12] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8, 2016.

[13] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. GEINet: View-invariant gait recognition using a convolutional neural network. In *2016 International Conference on Biometrics (ICB)*, pages 1–8, 2016.

[14] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 1 of *NIPS'14*, pages 568–576, Cambridge, MA, USA, 2014. MIT Press.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[16] A. Sokolova and A. Konushin. Pose-based deep gait recognition. *IET Biometrics*, 8:134–143, 2017.

[17] N. Takemura, Y. Makihara, and D. Muramatsu. On input/output architectures for convolutional neural network-based cross-view gait recognition. In *IEEE Trans Circuits Syst Video Technol 1–1*, 2017.

[18] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10, 12 2018.

[19] T. Whytock, A. Belyaev, and N. M. Robertson. Dynamic distance-based shape features for gait recognition. *Journal of Mathematical Imaging and Vision*, 50(3):314–326, Nov 2014.

[20] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39, 03 2016.

[21] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh. Gaitgan: Invariant gait feature extraction using generative adversarial networks. In *2017 IEEE Conference*

on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 532–539, 2017.

[22] S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang. Invariant feature extraction for gait recognition using only one uniform model. *Neurocomputing*, 239:81 – 93, 2017.

[23] S. Yu, D. Tan, and T. Tan. A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. In *Proc. of the 18-th International Conference on Pattern Recognition (ICPR)*, volume 4, pages 441–444, 2006.

[24] C. Zhang, W. Liu, H. Ma, and H. Fu. Siamese neural network based gait recognition for human identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2832–2836, Cham, 2016.

[25] X. Zhang, S. Sun, C. Li, X. Zhao, and Y. Hu. Deepgait: A learning deep convolutional representation for gait recognition. In *Biometric Recognition*, pages 447–456, Cham, 2017. Springer International Publishing.