МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра компьютерных технологий и систем

КОСТЮКЕВИЧ Полина Валерьевна

РАЗРАБОТКА СИСТЕМЫ РАСПОЗНАВАНИЯ ЭМОЦИЙ ЧЕЛОВЕКА НА ОСНОВЕ АНАЛИЗА ЕГО ДВИЖЕНИЙ В ВИДЕОПОТОКЕ

Дипломная работа

Научный руководитель:
Старший преподаватель кафедры КТС
В. А. Кулинкович

Допущена к защите
«__» _____ 2025 г.
Заведующий кафедрой компьютерных технологий и систем
доктор педагогических наук,
кандидат физико-математических наук,
профессор В.В. Казачёнок

Минск, 2025

ОГЛАВЛЕНИЕ

ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ

СКЛД — Система кодирования лицевых движений;

ANN (Artificial Neural Network) — искусственная нейронная сеть;

API (Application Programming Interface) — программный интерфейс приложения;

CNN (Convolutional Neural Network) — сверточная нейронная сеть;

CPU (Central Processing Unit) — центральный процессор;

GPU (Graphics Processing Unit) — графический процессор;

HOG (Histogram of Oriented Gradients) — гистограмма направленных градиентов;

LSTM (Long Short-Term Memory) — сеть с долгой краткосрочной памятью;

PAF (Part Affinity Fields) — поля аффинности частей тела, используемые в алгоритмах извлечения позы;

RGB (Red, Green, Blue) — цветовая модель, использующая трехканальное представление цвета (красный – зеленый – синий);

VGG-19 — архитектура сверточной нейронной сети Visual Geometry Group.

РЕФЕРАТ

Дипломная работа, 50 страниц, 9 рисунков, 5 таблиц, 24 формулы, 25 источников.

Ключевые слова: ВИДЕОПОТОК, ДЕТЕКЦИЯ ОБЪЕКТОВ, РАСПОЗНАВАНИЕ ЭМОЦИЙ, НЕЙРОННЫЕ СЕТИ, КЛАССИФИКАЦИЯ.

Объект исследования — алгоритмы и методы обнаружения ключевых точек лица и тела на изображениях и в видеопотоке.

Предмет исследования — применение инструментов распознавания человека в кадре для анализа и классификации невербальных проявлений эмоционального состояния в видеопотоке.

Цель исследования — анализ методов распознания эмоций человека на основе изменений в мимике и движениях тела, разработка системы распознавания эмоций в видеопотоке путем применения комбинированного подхода, сочетающего анализ мимики и положения тела.

Методы исследования: теоретические: анализ научной литературы, системный анализ, сравнительный анализ алгоритмов; практические: компьютерное моделирование, разработка алгоритма обработки видеопотока, тестирование системы.

В дипломной работе рассмотрены основные аспекты распознавания эмоций на основе анализа изображений и видео, изучены существующие инструменты и решения в данной предметной области. Исследованы подходы к классификации эмоций. Проведен сравнительный анализ алгоритмов детекции лиц и классификации эмоциональных состояний.

В результате разработано приложение, позволяющее определять эмоциональное состояние пользователя на основе изменения его мимики и положения тела в видеопотоке.

Области возможного практического применения: интерактивные развлекательные системы, системы человеко-компьютерного взаимодействия, системы безопасности и видеонаблюдения, психологические исследования и терапия, маркетинговые исследованиях.

.

РЭФЕРАТ

Дыпломная праца, 50 старонак, 9 малюнкаў, 5 табліц, 24 формулы, 25 крыніц.

Ключавыя словы: ВІДЭАПАТОК, ДЭТЭКЦЫЯ АБ'ЕКТАЎ, РАЗНАЗНАВАННЕ ЭМОЦЫЙ, НЕЙРОННЫЯ СЕТКІ, КЛАСІФІКАЦЫЯ.

Аб'ект даследавання — алгарытмы і метады выяўлення ключавых кропак твару і цела на малюнках і відэа.

Прадмет даследавання — прымяненне інструментаў распазнання чалавека ў кадры для аналізу і класіфікацыі невербальных праяў эмацыйнага стану ў відэа.

Мэта даследавання — аналіз метадаў распазнання эмоцый чалавека на аснове змен у міміцы і рухах цела, распрацоўка сістэмы распазнання эмоцый у відэаструмені шляхам ужывання камбінаванага падыходу, які спалучае аналіз мімікі і становішчы цела.

Метады даследавання: тэарэтычныя: аналіз навуковай літаратуры, сістэмны аналіз, параўнальны аналіз алгарытмаў; практычныя: камп'ютарнае мадэляванне, распрацоўка алгарытму апрацоўкі відэаструменю, тэсціраванне сістэмы.

У дыпломнай рабоце разгледжаны асноўныя аспекты распазнання эмоцый на аснове аналізу выяў і відэа, вывучаны існуючыя інструменты і рашэнні ў гэтай прадметнай галіне. Даследаваны падыходы да класіфікацыі эмоцый. Праведзены параўнальны аналіз алгарытмаў дэтэкцыі асоб і класіфікацыі эмацыйных станаў.

У выніку распрацавана праграма, якая дазваляе вызначаць эмацыйны стан карыстача на аснове змены яго мімікі і становішчы цела ў відэа.

Вобласці магчымага практычнага прымянення: інтэрактыўныя забаўляльныя сістэмы, сістэмы чалавека-кампутарнага ўзаемадзеяння, сістэмы бяспекі і відэаназірання, псіхалагічныя даследаванні і тэрапія, маркетынгавыя даследаванні.

ABSTRACT

Graduate work, 50 pages, 9 figures, 5 tables, 24 formulas, 25 sources.

Keywords: VIDEO STREAM, OBJECT DETECTION, EMOTION RECOGNITION, NEURAL NETWORKS, CLASSIFICATION.

The object of the research is algorithms and methods for detecting key points of the face and body in images and in a video stream.

The subject of the research is the use of human recognition tools in the frame for the analysis and classification of non-verbal manifestations of emotional state in a video stream.

The purpose of the research is to analyze methods for recognizing human emotions based on changes in facial expressions and body movements, and to develop a system for recognizing emotions in a video stream by using a combined approach that combines the analysis of facial expressions and body position.

Research methods: theoretical: analysis of scientific literature, systems analysis, comparative analysis of algorithms; practical: computer modeling, development of a video stream processing algorithm, system testing.

The thesis examines the main aspects of emotion recognition based on image and video analysis, and studies existing tools and solutions in this subject area. Approaches to emotion classification are investigated. A comparative analysis of face detection and emotional state classification algorithms is conducted.

The result of the work is an application that allows determining the user's emotional state based on changes in their facial expressions and body position in the video stream.

Areas of possible usage: interactive entertainment systems, human-computer interaction systems, security and video surveillance systems, psychological research and therapy, marketing research.

ВВЕДЕНИЕ

Распознавание эмоций относится к процессу идентификации и интерпретации эмоциональных состояний человека на основе анализа различных сигналов, таких как мимика, жесты, тон голоса и другие невербальные выражения. Этот процесс может осуществляться как человеком, так и автоматизированными системами, использующими технологии компьютерного зрения и машинного обучения. Его цель — дать представление об эмоциональном состоянии людей и дать возможность системам реагировать соответствующим образом.

Интерес к распознаванию эмоций на основе анализа мимических выражений можно объяснить рядом факторов. Прежде всего, изображения лиц легко доступны и широко распространены в повседневной жизни. Положение тела человека также является маркером проявления определенных эмоциональных состояний. Комбинация этих признаков имеет высокую информативность для распознавания эмоций, так как сокращения и растяжения мышц лица, вызывающие мимические изменения, вместе с другими физическими признаками могут явно выражать или намекать на определенные эмоции. Более того, сбор больших наборов фото и видеоданных значительно проще в сравнении со сбором аудиозаписей, требующих чистый звук, или данных о пульсе, давлении и других показателях, для измерения которых необходимо специальное оборудование.

Автоматическое распознавание эмоций имеет значительные преимущества по сравнению с анализом, проводимым человеком. Предварительно обученные модели способны обрабатывать большое количество изображений и видеопотоков за минимальное время. На текущий момент существует большое количество систем, ориентированных на анализ ключевых точек лица и последующее выявление закономерностей, которые имеют широкое применение в таких областях как медицина, безопасность, образование, телекоммуникации и маркетинг, что подчеркивает их важность и актуальность. Широкое развитие нейронных сетей, способных обрабатывать информацию о расположении человека в видеопотоке и выделять ключевые точки тела, открывает возможности для дальнейшего анализа признаков лица и позы человека в совокупности.

Цель данной дипломной работы заключается в проведении исследований по разработке системы автоматического распознавания эмоций человека путем применения комбинированного подхода, сочетающего анализ мимики и положения тела.

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

- Провести анализ предметной области;
- Рассмотреть психофизиологические аспекты восприятия эмоций;
- Проанализировать существующие методы и подходы к распознаванию эмоций;
- Изучить алгоритмы, используемые для извлечения признаков из видеопотока;
- Провести сравнительный анализ инструментов, предоставляющих возможность извлечения информации о позе человека из видеопотока;
- Разработать схему обработки видеопотока и анализа движений, включая этапы предварительной обработки, определения ключевых признаков и классификации;
- Протестировать разработанную систему;
- Провести анализ полученных результатов;
- Выявить практическую значимость применения разработанного алгоритма.

Работа состоит из трех глав: Понятие эмоций и их классификация в контексте распознавания; Методы и технологии распознавания движений для анализа эмоций; Проектирование и реализация системы распознавания эмоций.

В первой главе проводится анализ предметной области, рассматриваются особенности проявления эмоций в мимике, жестах и движениях человека, анализируются существующие подходы к определению и классификации эмоций.

Во второй главе рассматриваются методы компьютерной обработки, а также нейросетевые архитектуры, используемые в рамках задачи распознавания лица и позы человека на изображениях и кадрах видеопотока; проводится сравнительный анализ существующих решений, предоставляющих возможности детекции ключевых точек лица и тела человека.

Третья глава описывает особенности реализации системы, использующей MediaPipe, для комплексного извлечения признаков и последующей классификации эмоционального состояния на основе их анализа.

ГЛАВА 1
ПОНЯТИЕ ЭМОЦИЙ И ИХ КЛАССИФИКАЦИЯ В КОНТЕКСТЕ РАСПОЗНАВАНИЯ

1.1 Актуальные аспекты распознавания эмоций

Термином «эмоция» принято обозначать особую группу психических процессов и состояний, в которых выражается субъективное отношение человека к внешним и внутренним событиям [1, с. 504]. Эмоции играют ключевую роль в распознавании и интерпретации человеческого поведения.

Эмоции как состояния представляют собой относительно устойчивые переживания, которые длятся некоторое время (спокойствие, грусть, тревога) и влияют на общее самочувствие и поведение, формируя эмоциональный фон. Понятие эмоций как процесса подразумевает динамические переживания, связанные с конкретными событиями, которые имеют фазы возникновения, развития и затухания (испуг, гнев). Фаза возникновения определяется периодом после появления некоторого стимула — внешнего возбудителя или внутреннего состояния, и реакцией организма на него. Следующая фаза развития характеризуется физиологическими изменениями, которые проявляются в мимике, изменении голоса, телесных реакциях. Далее происходит затухание, спад эмоционального напряжения и возвращение организма в состояние равновесия.

Выражения лица представляют собой формы невербального общения, которые передают человеческие эмоции. Согласно мета-анализу, основанном на данных 127 исследований, с суммарным количеством участником более 45.000, диапазон влияния невербальных выражений на передачу и восприятие информации оценивается в 40-80% в зависимости от типа коммуникации, контекста и культурных норм [8].

В последние годы наблюдается значительный интерес к определению психоэмоционального состояния человека, что связано с его потенциальным влиянием на различные сферы жизни. Особенно актуально распознавание эмоций по изображениям лиц, поскольку лица легко доступны и часто встречаются в повседневной жизни. Мимика и выражения лиц могут явно указывать на эмоциональные состояния, что делает их ключевыми объектами для анализа.

Распознавание эмоций относится к аффективным вычислениям, которые исследуют взаимодействие человека и машины. Для успешного взаимодействия с пользователями крайне важно, чтобы машины могли точно понимать и классифицировать текущие эмоции человека. Это понимание

является основой для разработки систем, которые могут адаптироваться к эмоциональному состоянию пользователей и улучшать взаимодействие с ними. Например, в области здравоохранения такие системы могут помочь в диагностике и мониторинге психоэмоциональных состояний пациентов, что важно для оказания своевременной помощи. Кроме этого, большое значение подобные технологии могут иметь в таких областях как криминалистика, системы безопасности, маркетинговые исследования, online-обучение и др.

1.2 Подходы к классификации эмоций

Разные подходы к интерпретации эмоций приводят к многообразию способов их классификации. Выделяют два основных направления для систематизации множества эмоций: непрерывный и дискретный.

Непрерывная модель классификации предполагает рассмотрение эмоций как спектра, где они могут варьироваться по различным измерениям, таким как валентность и интенсивность. Данный подход был впервые предложен немецким ученым Вильгельмом Вундтом и позже развит американским психологом Гарольдом Шлосбергом, который выделил три измерения: «приятно-неприятно», «напряжение-расслабление» и «возбуждение-спокойствие» [20]. Позже Рассел в 1980 году предложил модель круга и предположил, что все эмоции можно расположить по окружности, управляемой двумя независимыми измерениями: гедоническим («удовольствие-неудовольствие») и возбуждение («активность-покой») [19].

Таким образом, непрерывные модели часто визуализируются в виде кругов или графиков, что позволяет увидеть, как разные эмоции расположены относительно друг друга. Непрерывная модель классификации эмоций представляет собой довольно гибкий и реалистичный подход к изучению и пониманию эмоций, однако в силу отсутствия четких границ между категориями, использование данного способа может привести затруднениям в классификации, особенно при наличии смешанных эмоциональных состояний. Алгоритмы машинного обучения, работающие с непрерывными данными, могут требовать больших объемов размеченных данных для точного обучения.

В отличие от непрерывного, дискретный подход к классификации эмоций предполагает, что существует ограниченное количество базовых эмоций, которые являются универсальными и биологически предопределенными. Идея строгого выделения базовых эмоций начала активно развиваться во времена древнегреческих философов. Например, Аристотель выделял 11 основных видов эмоций: страх, гнев, сострадание,

радость, ненависть, удивление, стыд, сожаление, гордость, доброта, надежда. Позже стоики, стремившиеся к рационализации мышления и чувств, выдвинули упрощенную систему, выделив всего 4 основные эмоции: печаль, страх, желание и радость.

В свое время Чарльз Дарвин написал книгу о выражении эмоций у человека, отметив, что мимически они проявляются одинаково у людей разных возрастов, рас и национальностей [11]. Эта идея легла в основу дальнейших исследований в области эмоционального восприятия, и в настоящее время наиболее известными представителями данного подхода являются Пол Экман и Уоллес Фризен. Они выделили 6 базовых эмоций на основе универсальности и биологической предопределенности: радость, удивление, гнев, отвращение, печаль и страх [13]. Система кодирования лицевых движений (СКЛД), основой которой являются перечисленные эмоции, будет подробнее рассмотрена далее.

Выбор дискретного подхода с базовыми эмоциями для решения задачи автоматического распознавания эмоций обоснован его простотой, универсальностью и практической применимостью, что делает его более подходящим для разработки эффективных систем.

1.3 Система кодирования лицевых движений

Одним из наиболее значимых и широко используемых инструментов классификации физического выражения эмоций является Система кодирования лицевых движений (СКЛД), или Facial Action Coding System (FACS), разработанная Полом Экманом и Уоллесом Фризеном в 1978 году. Эта система предоставляет методологию для детального анализа выражений лица, позволяя исследовать, как эмоциональные состояния отражаются в мимических изменениях [13].

Согласно данной системе, лицо можно разделить на три основные зоны: лоб – брови – глаза; нос – щеки; рот – подбородок. Как правило, выражение различных эмоций сопровождается изменения в каждой из этих зон. Таким образом, согласно подходу Экмана и Фризена выделяется шесть базовых эмоций для анализа:

— радость,
— удивление,
— гнев,
— отвращение,
— печаль
— страх.

Данный набор позволяет охватить наибольшее число реальных выражений лиц, так как другие эмоции являются естественным и более тонким развитием базовых.

В результате ряда исследований, Экман заключил, что базовые эмоции формируются благодаря нейронным программам, заложенным в человека с рождения. Это означает, что лицевые мышцы реагируют на эмоциональные состояния рефлекторно, и эти рефлексы являются аналогичными для всех людей, независимо от культурных различий. Таким образом, мимические изменения служат универсальными индикаторами эмоционального состояния.

В таблице 1.1 приведены примеры описания состояний, сопровождающихся базовыми эмоциями и типичными проявлениями в лицевых выражениях.

Таблица 1.1 – Описание базовых эмоций и их лицевых выражений

Эмоция	**Описание**	**Лицевые выражения**
Радость	Позитивное эмоциональное состояние, связанное с удовлетворением	Приподнятые уголки губ, улыбка, приподнятые щеки, расслабленные глаза
Удивление	Эмоция, возникающая в ответ на неожиданное событие или информацию	Открытый рот, приподнятые брови, широко раскрытые глаза
Страх	Эмоция, возникающая в ответ на угрозу или опасность	Расширенные глаза, приподнятые брови, открытый рот, напряженные мышцы лица
Отвращение	Эмоция, возникающая в ответ на неприятные объекты или ситуации	Сжатые губы, поднятый нос, морщины в области лба
Гнев	Эмоция, связанная с чувством недовольства или агрессии	Сжатые губы, нахмуренные брови, расширенные ноздри,
Печаль	Эмоция, связанная с разочарованием, потерей или горем	Опущенные уголки губ, слегка нахмуренные брови, опущенные веки

Данная система является мощным инструментом для изучения не только базовых эмоций, но и более сложных эмоциональных состояний, которые могут возникать в результате комбинации базовых эмоций. СКЛД позволяет вручную закодировать практически любое анатомически возможное выражение лица, конструируя его из отдельных действий, называемых

единицами действия, и определяя необходимое время для воспроизведения каждого выражения.

Система кодирования лицевых движений также находит применение в клинических исследованиях. Например, она используется для анализа степени депрессии и для измерения боли у пациентов, которые не могут описать свои ощущения словами. Это делает СКЛД важным инструментом в контексте психологии и медицины, где точное понимание эмоциональных состояний может иметь критическое значение для диагностики и лечения.

1.4 Обзор решений в области автоматического распознания эмоций по лицевым движениям

В данном разделе рассмотрены существующие реализации технологий распознавания эмоций из видео, которые имеют широкое распространение. В таблице 1.2 приведен ретроспективный анализ таких популярных инструментов как FaceReader, Affectiva, Microsoft Face API, далее приведено более подробное описание особенностей и возможностей каждого их них.

Таблица 1.2 – Сравнение популярных инструментов для распознавания эмоций

Критерий	**Affectiva**	**Microsoft Azure Face API**	**FaceReader**
Модель эмоций	6 базовых эмоций + 20 доп. показателей	6 базовых эмоций + презрение + нейтральное	6 базовых эмоций + нейтральное
Детекция точек	68 точек	468 точек	>500 точек
Детализация	Микровыражения, зрачки	Возможность определения презрения	Полный СКЛД-анализ
Архитектура	CNN + LSTM	ResNet-152 + Transformer	Active Appearance Model
Интеграция	SDK для мобильных устройств	REST API	Windows SDK + API
Требования	GPU	Достаточно CPU	Требуется камера с ИК-подсветкой

1.4.1 Affectiva

Affectiva – это одна из ведущих компаний, разрабатывающих решения для распознавания эмоций. Использует технологии распознавания лицевых выражений, жестов, языка тела путем анализа видео для оценки эмоций [4].

База данных, используемая в алгоритмах Affectiva, формируется на основе материалов, включающих видео людей, взаимодействующих с медиа и рекламой, записи водителей в реальных и смоделированных условиях, а также популярные GIF-изображения, отражающие универсальные и ситуативный эмоции.

На текущий момент архив насчитывает более 13 миллионов видео, снятых из порядка 90 стран, что предоставляет возможности для анализа выражений эмоций, учитывая культурный контекст. Такой анализ является одним из направлений инструментов, предоставляемых Affectiva, наряду с другими, имеющими применение в маркетинге, образовании, медицинских исследованиях. Продукт поддерживает как мобильные устройства, так и серверные решения. Affectiva использует искусственный интеллект и машинное обучение для анализа лицевых выражений и аудиофизических признаков, таких как тон и интонация речи. Данная система способна работать при разных условиях освещения, различных углах наклона головы и даже в условиях частичной скрытости лица.

Модули Affectiva включают Emotion SDK, реализующий распознавание эмоций по лицу в реальном времени; Affdex SDK, ориентированный на анализ мимики, взгляда, возраста, этнической принадлежности; Affectiva Media Analytics, активно используемый для анализа предварительно записанных видео, таких как рекламные ролики или кино. Решения Affectiva позволяются распознавать более базовые эмоции, показывая точность 70-80% в зависимости от ракурса, освещения, качества видео.

В 2021 году шведская компания Smart Eye, которая является лидером в системах отслеживания взгляда, приобрела Affectiva и разработка сместилась в сторону алгоритмов, ориентированных на автономные автомобили и безопасность. Например, Volvo, BMW активно используют эти технологии для определения психоэмоционального состояния водителя с целью обнаружения усталости и других уязвимых состояний для предотвращения аварийных ситуаций.

1.4.2 Microsoft Azure Face API

Microsoft Azure Face API входит в состав инструментов Azure AI services и представляет собой облачный сервис, предоставляющий возможности для анализа и распознавания лиц на изображениях и видео [7]. Среди множества функций, предоставляемых Face API, одной из ключевых является распознавание эмоций. Этот компонент анализирует выражения лиц и определяет различные эмоциональные состояния человека на основе

визуальных признаков. Идентифицирует 6 базовых эмоциональных состояний, согласно классификации Экмана, а также нейтральное.

В основе Azure Face API лежит гибридная архитектура, имеющая модуль детекции лиц, построенный на модифицированной версии ResNet, оптимизированной для работы с изображениями высокого разрешения.

Благодаря использованию современных методов машинного обучения Face API достигает высокой точности в распознавании эмоций в реальных условиях. Алгоритмы могут распознавать эмоции, даже если лицо слегка повернуто или частично скрыто. Результат обнаружения лиц программа возвращает в виде координат прямоугольника, ограничивающего лицо.

Кроме распознавания эмоций, Face API предлагает дополнительные функции, такие как идентификация лиц, проверка личностей и анализ демографических данных (возраст, пол и т. д.). Эти возможности позволяют интегрировать Face API в широкий спектр приложений, от систем безопасности до приложений в области маркетинга и персонализированного обслуживания клиентов.

1.4.3 FaceReader (Noldus Information Technology)

FaceReader — это высокотехнологичное программное обеспечение, разработанное компанией Noldus Information Technology для анализа выражений лиц и распознавания эмоций. Оно активно используется в научных исследованиях, маркетинге, психологии и других областях, где необходимо изучать и интерпретировать эмоциональное состояние людей.

FaceReader способен точно выявлять степень выраженности каждой эмоции, предоставляя подробную информацию о том, насколько интенсивно выражена каждая эмоциональная реакция. Это позволяет исследователям и специалистам более глубоко понимать эмоциональные реакции пользователей и принимать обоснованные решения на основе полученных данных.

Одной из ключевых особенностей FaceReader является поддержка анализа видеопотоков в реальном времени, что делает его эффективным инструментом для мониторинга реакций в условиях живого общения, таких как фокус-группы или интервью. Кроме того, FaceReader может использоваться для анализа записанных видео, что позволяет проводить ретроспективные исследования.

Данный продукт занимает нишу в академической и прикладной психофизиологии и используется чаще всего в рамках научных исследований, по этой же причине не поддерживает мобильные платформы.

В основе FaceReader лежит гибридная система анализа, комбинирующая модель для трекинга более 500 ключевых точек лица, глубокие нейронные сети для классификации шести базовых эмоций, предложенных Экманом.

1.5 Связь положения тела с эмоциональным состоянием

Согласно ряду исследований, эмоциональное состояние характеризуется изменениями не только в мимике, но и в жестах, движениях человека в целом. Причем чем сильнее выражается эмоция, тем более выражен язык тела человека [6]. Самые явные закономерности, характерные для многих людей: опущенная голова при выражении грусти или стыда, опора головы на ладонь или кулак при скуке.

Общую информацию о взаимосвязи выражения эмоции и движений тела, основанную на исследованиях Дарвина, Экмана, а также Уолботта, автора работы «Телесное выражение эмоций» [23], можно заключить в таблицу 1.3:

Таблица 1.3 – Описание базовых эмоций и их лицевых выражений

Эмоция	**Характерные движения и позы**
Радость	Бесцельные движения, хлопки, подпрыгивания; тело держится прямо, голова поднята
Удивление	Приподнятые плечи, руки поднесены к лицу; Голова откидывается назад или слегка наклоняется вперед
Страх	Голова располагается ближе к плечам, «втягивается»; плечи подняты, руки прижаты к бокам или к груди, рука может перекрывать лицо
Отвращение	Движение руками перед собой или руки прижаты к бокам, плечи подняты; голова отвернута, корпус отдален от объекта
Гнев	Жесты становятся беспорядочными и резкими; человек может ходить вперед-назад, сжимать кулаки, плечи расправлены
Печаль	Неподвижность, пассивность; голова опущена

1.6 Заключение по первой главе

Разные подходы к классификации эмоций отражают сложность и многообразие человеческих чувств. Дискретные модели, такие как модель Экмана, помогают точно идентифицировать эмоции через мимическое выражение, в то время как непрерывные модели предлагают более гибкий и реалистичный взгляд на эмоциональные состояния, учитывая их динамику и контекстуальные изменения.

Система кодирования лицевых движений является мощным инструментом для распознавания эмоций, что делает ее крайне актуальной для использования при разработке систем, способных интерпретировать эмоциональные состояния на основе видеоанализа. Успешное распознавание эмоций может значительно повысить эффективность различных программ, начиная от личных помощников и заканчивая терапевтическими приложениями.

Современные модели машинного обучения для распознавания эмоций находят широкое применение в различных сервисах и приложениях. Эти технологии позволяют с высокой точностью анализировать психоэмоциональное состояние пользователей по фото- и видеоданным, открывая новые возможности в таких областях, как цифровая психометрия, адаптивные интерфейсы и телемедицина. Особую ценность представляют решения, способные работать в реальном времени и адаптироваться к индивидуальным особенностям пользователей.

ГЛАВА 2
МЕТОДЫ И ТЕХНОЛОГИИ АНАЛИЗА ДВИЖЕНИЙ ДЛЯ РАСПОЗНАВАНИЯ ЭМОЦИЙ

2.1 Анализ исходных данных

Исходными данными для компьютерной обработки с целью последующего извлечения признаков являются изображения, представляющие собой видеокадры, в которых зафиксированы мимические изменения человеческих лиц и положений тела. В рамках данной работы будет использоваться открытый набор видеозаписей, включающий различные эмоциональные состояния.

Допускаются незначительные дефекты в видеокадрах, такие как шум, низкий контраст или нечеткость, которые могут возникнуть в процессе записи. Эти проблемы будут устранены в ходе предварительной обработки входных изображений, что позволит повысить точность алгоритма распознавания эмоций.

2.2 Этапы автоматического распознавания эмоций

В контексте данной задачи видео понимается как предварительно обработанная временная последовательность кадров, где каждый кадр обозначается согласно формуле (2.1):

$$V \in R^{T \times H \times W}, \tag{2.1}$$

где *T, H, W* представляют собой количество кадров, высоту и ширину кадра соответственно. Каждый кадр разделяется на $N = \frac{H \times W}{P^2}$ равных участков, каждый из которых можно анализировать отдельно.

К каждому участку добавляются временные векторы, что позволяет обеспечить последовательную классификацию эмоций на основе каждого из этих векторов, используя предыдущее состояние. Текущая информация затем пересчитывается в конечные значения для формирования окончательной предсказательной информации, которая передается классификатору.

Как правило, процесс распознавания эмоционального состояния по лицевым признакам из видеопотока можно разделить на следующие этапы:

1. Регистрация кадров видеопотока;
2. Предварительная обработка кадров;

3. Выделение области лица на изображениях;
4. Определение основных зон лица, выделение ключевых точек и анализ их расположения;
5. Классификация эмоций на основе значений извлеченных ключевых признаков.

Сам алгоритм можно описать следующей схемой (Рисунок 2.1).

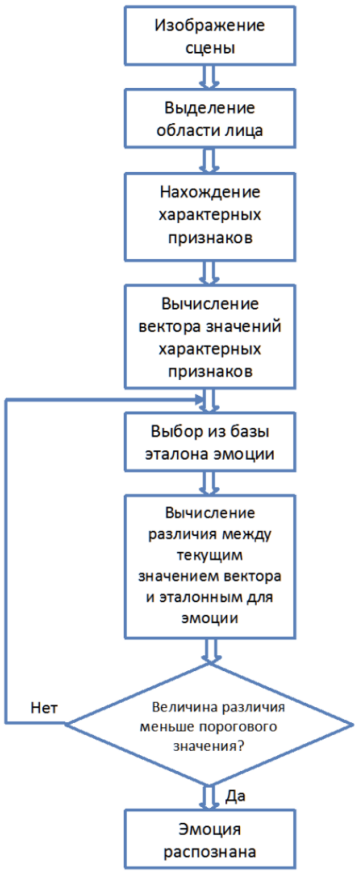


Рисунок 2.1 – Алгоритм распознавания эмоций по лицевым признакам

2.3 Предварительная обработка

Набор данных обычно предварительно обрабатывается, чтобы обеспечить согласованность данных и устранить шум [2]. В системах анализа и распознавания изображений при предварительной обработке изображения используется процедура преобразования цветного изображения в полутоновое. Суть метода заключается в изменении градации каждого пикселя изображения в соответствии с определенным правилом. В результате каждый пиксель имеет 1 канал (от 0 до 255 уровней серого) вместо 3 каналов для цветного изображения (RGB или YCrCb).

Существует несколько методов преобразования модели RGB в оттенки серого. Самый простой метод заключается в том, чтобы значению серого присвоить среднее значение компонентов с наибольшим и наименьшим значением среди красного (R), зеленого (G) и синего (B) цветов в обрабатываемой точке (2.2):

$$grayscale = \frac{max(R, G, B) + min(R, G, B)}{2}. \quad (2.2)$$

Явным недостатком использования данного алгоритма является то, что один из компонентов RGB не используется.

Более оптимизированным является метод усреднения, в котором вычисляется среднее значение все трех компонентов (красного, зеленого и синего) (2.3):

$$grayscale = \frac{R + G + B}{3}. \quad (2.3)$$

По причине того, что человеческий глаз наиболее чувствителен к зеленому, затем к красному и после к синему, в результате глубокого анализа было выведено следующее уравнение представления значения серого из компонентов цветовой модели RGB (2.4):

$$grayscale = Y = 0{,}3R + 0{,}59G + 0{,}11B. \quad (2.4)$$

Этот шаг позволяет оптимизировать общий внешний вид изображения и облегчает дальнейшую обработку.

2.4 Методы компьютерного зрения для извлечения признаков

Выделение области, в которой располагается человек, является ключевым этапом в процессе распознавания эмоций. Следующие процедуры обработки возможны после выделения области лица и тела, поскольку связаны с определением размеров этой области, ее конфигурации. При распознавании лиц в большинстве случаев происходит сравнение биометрических шаблонов (векторов признаков, характеризующих цифровое изображение лица).

2.4.1 Метод Виолы-Джонса

Для обнаружения объектов часто используется метод Виолы-Джонса, основанный на использовании каскадов классификаторов и интегральных изображений. Данный метод является фундаментальной основой для современных алгоритмов детекции человека в кадре, включая более сложные методы компьютерного зрения и нейросетевые подходы. Метод Виолы-Джонса сочетает высокую скорость обработки и минимальное количество ложных распознаваний.

Принцип работы алгоритма для распознавания объектов на изображении основан на технологии скользящего окна. Анализируемое изображение последовательно обрабатывается прямоугольной областью фиксированного размера, называемую «окном», которая перемещается по всей области кадра. Для каждого положения окна применяется серия классификаторов, определяющих наличие искомого объекта.

На основе входного изображения *I* создается интегральное изображение, основанное на суммировании пикселей в заданной области, для ускорения последующих вычислений. Для пикселя *(x, y)* интегральное изображение *L(x, y)* определяется как сумма по всем пикселям, находящимся выше и левее *(x, y)*, которая описывается формулой (2.5):

$$L(x,y) = \sum_{i=0}^{x}\sum_{j=0}^{y} I(i,j), 0 \le x \le N, 0 \le y \le M, \quad (2.5)$$

где $I(i,j)$ — значение яркости пикселя в координатах (i,j) исходного изображения,

N, M — размеры изображения по ширине и высоте.

Для оптимизации вычислений используется рекуррентная формула (2.6), позволяющая вычислить интегральное изображение за один проход, исключая дублирование вычислений.

$$L(x,y) = I(x,y) - L(x-1,y-1) + L(x,y-1) + L(x-1,y). \quad (2.6)$$

Следующий этап – применение признаков Хаара – специальных шаблонов, состоящих из смежных прямоугольных областей. Признаком называют отображение вида (2.7):

$$f: X \to D_f, \quad (2.7)$$

где D_f — множество допустимых значений признака.

Если заданы признаки $f_1, f_2, \dots, f_n$, то вектор признаков $x = (f_1(x), f_2(x), \dots, f_n(x))$ называется признаковым описанием объекта $x \in X$.

Признаковые описания допустимо отождествлять с самими объектами. Примеры примитивов признаков Хаара изображены на рисунке 2.2.

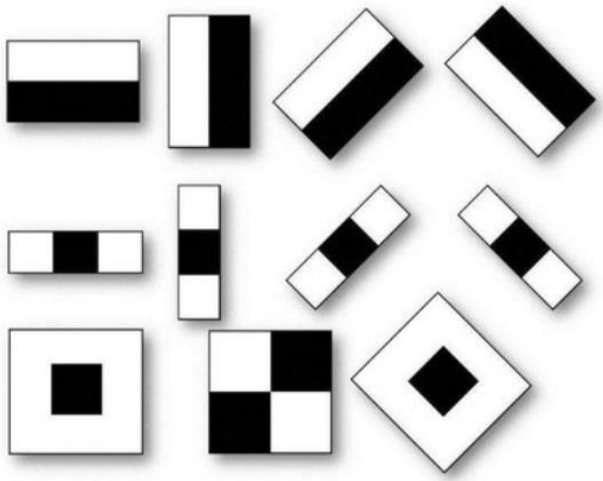


Рисунок 2.2 – Стандартные примитивы признаков Хаара

Для отбора оптимальных признаков на участках изображения используется метод бустинга. Этот метод работает по принципу последовательного добавления классификаторов, то есть каждый следующий алгоритм фокусируется на исправлении ошибок, допущенных предыдущим, тем самым создавая более точную модель за счет итеративного улучшения предсказаний.

Каждый признак используется для вычисления разности сумм яркостей между светлыми и темными участками согласно формуле (2.8):

$$F = X - Y, \tag{2.8}$$

где X — суммарная яркость пикселей в светлой области признака, Y — суммарная яркость пикселей в темной области признака в интегральном представлении изображения. Используя интегральное изображение, можно вычислять суммы за константное времени, что значительно ускоряет процесс детекции по сравнению с обычными методами, требующими *O(N)* времени для каждой области. Это делает алгоритм Виолы-Джонса эффективным и быстрым для задачи детекции лиц. Извлеченные области далее можно использовать для решения задачи классификации.

2.4.2 Гистограмма направленных градиентов

Гистограмма направленных градиентов (HOG) — это метод выделения признаков, основанный на анализе распределения градиентов яркости. Суть алгоритма детекции объектов данным способом основывается на идее, что форма объекта определяется не абсолютными значениями пикселей, а локальными изменениями яркости и их ориентацией в пространстве [10].

Для каждого пикселя предварительно подготовленного и нормализованного изображения вычисляются горизонтальный и вертикальный градиенты по формулам (2.9) и (2.10) соответственно:

$$G_x = I(x+1, y) - I(x-1, y), \tag{2.9}$$

$$G_y = I(x, y+1) - I(x, y-1), \tag{2.10}$$

где *I(x ,y)* — значение яркости пикселя с координатами *(x, y).*

Далее по формулам (2.11), (2.12) вычисляются величина и направление градиента:

$$|G| = \sqrt{G_x^2 + G_y^2}, \tag{2.11}$$

$$\theta = \arctan\left(\frac{G_y}{G_x}\right) \in [0, 180°]. \tag{2.12}$$

Изображение делится на ячейки небольшого размера, для каждой из которых строится гистограмма. Для каждого блока вычисляется гистограмма направленных градиентов, группируя направления по B интервалам (например, 9 интервалов по 20°). Вычисления описаны формулой (2.13):

$$H(b) = \sum_{(x,y)\in B} G(x,y) \cdot \delta\big(\theta(x,y)\big) \in b), \tag{2.13}$$

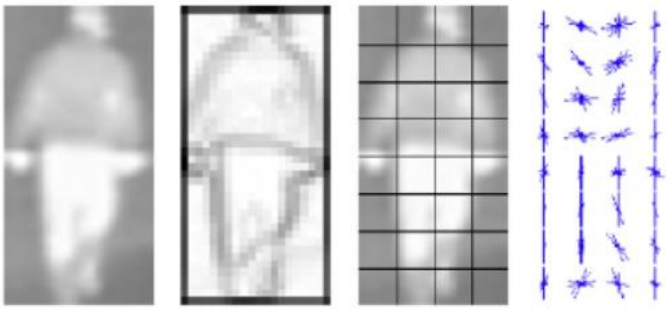
где $G(x, y)$ — величина градиента в точке (x, y),
δ — функция, которая равна 1, если направление градиента попадает в интервал b.

Далее значения гистограммы для каждого блока нормализуются согласно формуле (2.14):

$$H_{norm} = \frac{H(b)}{\sqrt{\sum H(b)^2 + \varepsilon^2}}, \tag{2.14}$$

где ε — некоторое малое значение.

Полученные гистограммы объединяются в один вектор признаков, который в дальнейшем может использовать для решения задачи классификации. Таким образом, формируется HOG-дескриптор, то есть описание ключевых особенностей изображения в более компактной форме, отражающей направления градиентов. При применении к изображениям, на которых расположен силуэт человека, HOG-дескриптор может иметь вид, подобный представленному на рисунке 2.3.



а б в г

Рисунок 2.3 – Пример изображения на разных этапах построения гистограммы направленных градиентов: *а* – исходное изображение, *б* – изображение градиента, *в* – изображение, разделенное на ячейки, *г* – HOG-дескриптор

2.5 Нейросетевые подходы для извлечения признаков

Современные системы анализа эмоций и движений человека основаны на глубоком обучении. Нейронные сети способны автоматически извлекать высокоуровневые признаки из видеопотока, такие ключевые точки лица и тела с учетом как пространственных, так и временных зависимостей.

Основные подходы включают:

— Сверточные нейронные сети (CNN) для детекции статических поз и мимики.

— Рекуррентные архитектуры (LSTM, GRU) для анализа динамики движений.

— Трансформеры для обработки долгосрочных зависимостей в видео.

Эти методы обеспечивают высокую точность распознавания даже в сложных условиях: при изменениях освещения и нестандартных ракурсах. В данном разделе рассматриваются основные архитектуры, их преимущества и недостатки в контексте распознавания эмоций через позу и мимику.

Сверточные нейронные сети (CNN) представляют собой архитектуру сетей, которая активно применяется для работы с изображениями. Такие сети

устойчивы к искажениям и изменению масштаба входных изображений, что делает их более эффективными при анализе изображений. На предварительно обработанное изображение накладывается фильтр, который применяется последовательно ко всей области изображения. Таким образом, формируются карты сверточных слоев. Так как ядро свертки для каждой карты признаков одно, то это позволяет нейронной сети научиться выделять признаки вне зависимости от их расположения во входном изображении и также приводит к значительному уменьшению параметров.

В задачах анализа динамики движений значительную роль играет способность модели учитывать временные зависимости в последовательностях данных. Традиционные сверточные нейронные сети эффективно работают с пространственными признаками, однако для обработки видеопотока, требуются архитектуры, способные запоминать и интерпретировать долгосрочные зависимости. В таком случае более актуальными являются рекуррентные сети, в частности LSTM (Long Short-Term Memory) и GRU (Gated Recurrent Unit). Основная проблема стандартных рекуррентных нейронных сетей заключается в затухании градиента, что затрудняет обучение на длинных последовательностях. LSTM и GRU решают эту проблему путем введения механизмов управляемых фильтров, которые позволяют регулировать поток информации.

В LSTM используются три типа фильтров, которые регулируют поступление новой информации, управление старой информацией и контролируют выходное значение. Например, при анализе видеопоследовательности LSTM может эффективно отслеживать изменения выражения лица, игнорируя кратковременные помехи. Однако такая архитектура обладает высокой вычислительной сложностью из-за большого числа параметров, что может приводить к замедленному обучению и риску переобучения на ограниченных данных.

GRU является упрощенной версией LSTM, объединяя фильтры забывания и входа в один механизм обновления. Это сокращает вычислительные затраты примерно на 30% без существенной потери точности в большинстве задач. GRU демонстрирует особенно хорошие результаты при работе с последовательностями средней длины, такими как анализ коротких видеофрагментов или жестов в режиме реального времени. Однако при обработке очень длинных зависимостей (например, трекинг позы в протяженных видео) GRU может уступать LSTM из-за менее гибкого управления памятью.

В практических реализациях часто используют гибридные подходы, сочетающие CNN для извлечения пространственных признаков и LSTM/GRU

для анализа временной динамики, что позволяет достичь баланса между точностью и вычислительной эффективностью.

2.6 Методы детекции и классификации эмоций

2.6.1 Одноступенчатые методы детекции

Одноступенчатые детекторы, такие как YOLO (You Only Look Once) или SSD (Single Shot MultiBox Detector), выполняют обнаружение лиц за один проход через сеть. Одноэтапные алгоритмы подразумевают использование одной глубокой конволюционной нейронной сети, выходом которой являются непосредственно искомые прямоугольники и метки классов.

Алгоритм работы YOLO можно описать следующим образом: изначально изображение приводится к определенному размеру, затем разбивается на блоки, формируемые сеткой размера $N \times N$. Каждый блок отвечает за предсказание объектов, центр которых попадает в данный блок, и содержит информацию о координатах центра ячейки *(x, y),* ширине и высоте *(w, h),* показателя уверенности в том, что объект находится в предполагаемой области (2.15).

$$\begin{cases} x = \sigma(t_x) + c_x, \\ y = \sigma(t_y) + c_y, \\ w = p_w e^{t_w}, \\ h = p_h e^{t_h}, \\ confidence = \sigma(t_c), \end{cases} \tag{2.15}$$

где σ — функция активации (сигмоида),
$t_w, t_h.$ — значения смещений
(c_x, c_y) — координаты ячейки,
(p_w, p_h) — размеры блока.

В результате имеется набор ограничивающих рамок. Так как с высокой вероятностью распознаваемый объект на изображении имеет размер, больший чем величина блока сетки, возникают избыточные ограничивающие рамки, соответствующие одному объекту. Для удаления дубликатов предсказаний используется метод non-maximum suppression (NMS) [21]. Если несколько ячеек предсказывают один и тот же объект, оставляется только предсказание с наивысшей уверенностью. Точности локализации оценивается с помощью формулы IoU (Intersection over Union — пересечение по объединению) (2.16):

$$IoU = \frac{A \cap B}{A \cup B}, \tag{2.16}$$

где A – предсказанная ограничивающая рамка,
B – истинная ограничивающая рамка.

Стандартным является значение порога $k = 0.5$. Если $IoU > k$, то результат считается приемлемым, $IoU=1$ означает полное совпадение. Если же значение ниже порога, то пересечения нет.

Наиболее подходящей для задач распознавания лица и позы человека является YOLOv8, где внедрены модуль Task Aligned Assigner для оптимального предсказания при перекрытии объектов, возможность выделения наклонных областей и отдельный режим Pose Estimation с 17 ключевыми точками.

SSD (Single Shot MultiBox Detector) также использует единую сеть, но в отличие от YOLO применяется несколько карт признаков на разных уровнях для предсказания объектов. SSD аналогично разбивает входное изображение на сетку размером $N \times N$, где каждая ячейка отвечает за предсказание объектов, чьи центры попадают в эту ячейку. Для каждой ячейки i и каждой рамки j сеть предсказывает поправки $\Delta = (\Delta_x, \Delta_y, \Delta_w, \Delta_h)$. Итоговая ограничивающая рамка вычисляется по формулам (2.17):

$$\begin{cases} x = d_x + \Delta_x d_w, \\ y = d_y + \Delta_y d_h, \\ w = d_w e^{\Delta_w}, \\ h = d_h e^{\Delta_h}, \end{cases} \tag{2.17}$$

где (d_x, d_y) — координаты ячейки,
(d_w, d_h) — размеры блока.

Преимущества одноступенчатых детекторов заключаются в их высокой скорости и простоте реализации. Несмотря на это, одним из главных недостатков является их точность, особенно в таких условиях как плохое освещение, различные углы обзора или наличие препятствий.

2.6.2 Многоступенчатые методы детекции

Многоступенчатые детекторы, такие как R-CNN (Region-based Convolutional Neural Networks), работают в несколько этапов. Сначала они выделяют области, которые могут содержать искомый объект, а затем применяют более сложные модели для уточнения результатов.

Сначала входное изображение подается в сверточную нейронную сеть, которая формирует карты признаков, отражающие ключевые особенности изображения на разных уровнях абстракции. Затем отдельный модуль, такой как Selective Search в оригинальном R-CNN или Region Proposal Network (RPN) в Faster R-CNN, генерирует набор регионов-кандидатов, потенциально содержащих объекты. Эти регионы, представляющие собой ограничивающие прямоугольники, далее подвергаются процедуре RoI pooling (Region of Interest pooling), которая приводит их к фиксированному размеру для последующей обработки.

На заключительном этапе каждый регион анализируется отдельно с помощью дополнительной CNN, которая выполняет две основные задачи: классификацию объекта внутри региона (например, определение, что это лицо) и уточнение координат ограничивающего прямоугольника. Такой подход обеспечивает высокую точность детекции, особенно в сложных сценах с перекрытиями объектов и изменяющимся освещением.

Однако многоступенчатые детекторы обладают двумя принципиальными ограничениями. Во-первых, они анализируют не все изображение целиком, а лишь отдельные регионы-кандидаты, что может приводить к потере контекстной информации, важной для корректного распознавания объектов в сложных сценах. Например, при детекции эмоций по лицу глобальный контекст (положение головы, окружающие объекты) может содержать полезные подсказки, которые игнорируются при поточечном анализе регионов.

Во-вторых, последовательная архитектура обработки делает эти методы вычислительно затратными и относительно медленными по сравнению с одноступенчатыми детекторами (например, YOLO или SSD). Необходимость отдельно обрабатывать каждый регион-кандидат приводит к дублированию вычислений и увеличивает время обработки, что особенно проблематично в задачах реального времени, таких как анализ видеопотоков или интерактивные системы.

2.7 Инструменты трекинга и анализа позы

Современные методы анализа движений тела основаны на алгоритмах Pose Estimation (оценка позы). Под позой в компьютерном зрении подразумевается набор ключевых точек и отрезков, соединяющих их, которые в результате формируют подобие скелета человека.

Существует два основных подхода к анализу положения частей тела человека в пространстве: маркерный и безмаркерный.

Первый способ основан на использовании специальных маркеров, закрепленных на ключевых анатомических точках и последующем определением их координат в пространстве. Процесс обработки информации о положении тела при использовании такого подхода можно разделить на три этапа: трекинг маркеров, фильтрация шумов с целью устранения дрожания и реконструкция скелета.

Безмаркерные методы анализа тела основаны на алгоритмах компьютерного зрения и глубокого обучения, которые определяют позу, движения и даже эмоции человека без использования физических маркеров или специального оборудования. Существует два направления, описывающие безмаркерные методы: «сверху-вниз» и «снизу-вверх». При анализе «сверху-вниз» сначала определяется область, в которой располагается человек, а затем выделяются части тела с использованием модели ключевых точек. Этот подход обеспечивает высокую точность в случаях, когда человек в кадре один, но его производительность существенно снижается при увеличении количества человек из-за необходимости последовательной обработки каждой обнаруженной области.

Алгоритм «снизу-вверх» в свою очередь подразумевает первоначально поиск и распознавание частей тела, а затем их группировку, определяя, какие из них принадлежат одному человеку [3]. Несмотря на то, что методы данной группы сложнее в реализации, так как требуют решения задача группировки точек, они демонстрируют лучшую производительность в сценах с множеством людей.

Рассмотрим подробнее современные инструменты, которые активно используются для решения задач распознания людей по фото и видео.

2.7.1 OpenPose

OpenPose — это библиотека, предоставляющая современную систему оценки позы человека. Данная система была разработана исследователями из Университета Карнеги-Меллона. Реализует комплексный подход к одновременной детекции ключевых точек тела, лица, кистей рук, стоп одного или нескольких людей в кадре в режиме реального времени.

Алгоритм, используемый OpenPose, относится к группе методов «снизу-вверх», то есть изначально определяются ключевые точки, которые далее группируются, формируя части тела. Этот подход показывает высокую эффективность при распознавании групп людей в кадре. Таким образом, реализуется возможность распознавания до 25 человек без потери точности.

В основе архитектуры OpenPose лежит двухэтапная нейросетевая модель [9], сочетающая сверточные нейронные сети с механизмом Part Affinity

Fields (PAF) — это векторное поле, которое кодирует направление и степень связи между парными точками тела, например, плечо-локоть или бедро-колено. Таким образом, система принимает RGB-изображение, генерирует тепловые карты (heatmaps), которые определяют вероятностное расположение ключевых точек. Области изображения с максимальными значениями на карте характеризуются наибольшей вероятностью присутствия ключевых точек. Далее с помощью PAF устанавливаются пространственные связи между обнаруженными точки, что позволяет корректно ассоциировать точки с людьми в кадре даже в условиях частичных перекрытий.

Более подробно архитектура сети представлена на рисунке 2.4.

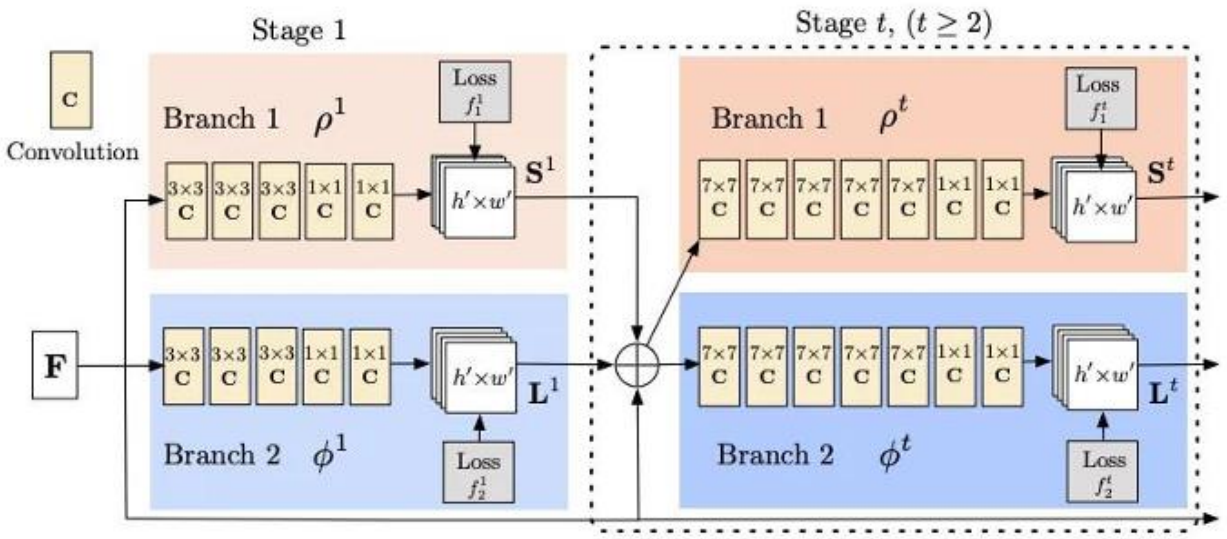


Рисунок 2.4 – Архитектура многоступенчатой CNN, используемой в реализации OpenPose

Первая часть сети представляет собой модифицированную архитектуру предварительно обученной нейронной сети VGG-19, где используются ядра 3×3 с шагом 1 и дополнением для сохранения пространственных размеров. Начальные слои содержат 64 карты признаков с последующим увеличением глубины до 512 каналов через серию сверточных слоев и операций макс-пулинга. Далее сеть разделяется на два параллельных потока, каждый из которых проходит через серию из 6 уточняющих стадий (refinement stages), где на каждой стадии используются свертки с ядрами 7×7 и 3×3, объединяющие предыдущие поля сходств, карты достоверности и исходные признаки. На последнем этапе увеличивается размерность карт признаков до исходного размера, что позволяет точно локализовать ключевые точки.

Процесс обработки изображения нейронной сетью для выделения ключевых точек имеет константную сложность вне зависимости от количества человек. Время анализа и группировки данных на следующем этапе имеет

сложность $O(n^2)$, где n – количество человек в кадре. Однако это время обработки не оказывает существенного влияния на общее время работы системы, так как оно значительно меньше времени выполнения CNN-обработки. Таким образом, OpenPose позиционируется как инструмент, ориентированный на способность эффективно распознавать несколько человек, при этом сохраняя скорость работы. Тем не менее, алгоритм имеет ряд случаев, в которых может происходить ложное или некорректное распознавание. Среди них сцены, где части тела одного человека частично перекрываются другим человеком или объектом; нестандартные позы, преимущественно те, где есть скручивания или поворот верхней части туловища.

2.7.2 MediaPipe

MediaPipe — кроссплатформенная библиотека от Google, специализирующаяся на трекинге мимики, жестов и позы в режиме реального времени. Активно используется для создания интерактивных игр с дополненной реальностью и масок лица в социальных сетях, а также в фитнес-приложениях для анализа правильности выполнения упражнений. Также применяется для мониторинга пациентов, включая диагностику заболеваний опорно-двигательного аппарата.

Google активно использует MediaPipe, интегрируя в свои сервисы:

— Google фото;
— Cloud Vision API;
— умные камеры NestCam;
— обнаружение объектов с помощью Google Lens;
— интерактивные рекламные форматы с отслеживанием жестов пользователей и т.д.

Преимуществом MediaPipe является поддержка устройств разного уровня. Система использует три уровня оптимизации: базовый CPU, обеспечивающий работу на устройствах без GPU и специализированных ускорителей; основной режим, использующий GPU-ускорение, поддерживает более плавный трекинг объектов в реальном времени с высокой производительностью в мобильных и десктопных приложениях; специализированные ускорители (Neural Processing Unit, Google Coral Edge TPU).

Архитектурной особенностью системы является модульный подход, основанный на концепции графов обработки медиаданных (Рисунок 2.5). В такой архитектуре обработка информации строится как последовательность взаимосвязанных узлов, называемых калькуляторами (MediaPipe Calculators),

каждый из которых выполняет определенную функцию: детекцию объектов, извлечение признаков, фильтрацию шумов или визуализацию результатов. Калькуляторы представляют собой специализированные модули, реализованные на C++, взаимодействующие между собой с помощью потоков данных. Пакеты данных поступают и выходят через порты в калькуляторе, которые задаются при конфигурации графа. При инициализации калькулятор объявляет тип данных, которые он способен принимать и генерировать. Это обеспечивает строгую типизацию и предотвращает ошибки на этапе исполнения. Каждый калькулятор реализует три ключевых метода жизненного цикла: open, process, close, которые вызываются соответственно при запуске графа, обработке данных и завершения всей обработки.

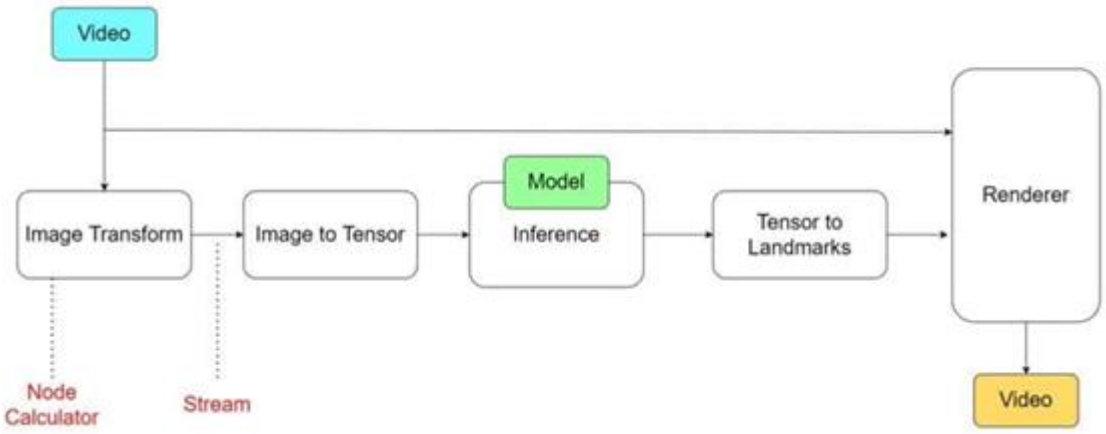


Рисунок 2.5 – Графы MediaPipe

Для трекинга позы человека используется каскадная архитектура BlazePose, которая состоит из детектора на основе легковесной CNN и модели предсказания ключевых точек. MediaPipe, в отличие от OpenPose реализует алгоритм «сверху вниз»: сначала обнаруживается область интереса, затем в этой области определяются ключевые точки тела. Система использует граф, включающий следующие компоненты: предобработку изображения, включающую нормализацию и изменение размера, применение ML-моделей, постобработку и визуализацию. Нейронная сеть работает со входным изображением размером 256×256 пикселей (RGB). Первый этап обработки включает модифицированную сверточную сеть с ядрами 3×3 и 5×5, который уменьшает размерность изображения до 32×32 пикселей при увеличении глубины признаков до 128 каналов. Далее сеть разделяется на два параллельных потока: первый поток генерирует heatmap-карты для 39 ключевых точек (33 точки тела, 4 точки лица и 2 точки для ориентации), а второй поток предсказывает 3D-координаты этих точек относительно корневой точки. Каждый поток состоит из последовательности сверточных

слоев с ядрами 3×3 и 1×1, которые сохраняют пространственную размерность 32×32. Для повышения точности MediaPipe использует механизм уточнения ключевых точек через дополнительный блок GRU обрабатывающий временные последовательности в видео. Финальный этап включает постобработку: устранение дубликатов точек и алгоритм трекинга для стабилизации результатов между кадрами. Для сглаживания результатов между кадрами используется оптический поток и фильтр Калмана — алгоритм оценки состояния системы, который предсказывает позицию точки, даже если часть тела, к которой она относится, временно скрыта.

В отличие от классических CNN, MediaPipe сочетает сверточные слои с рекуррентными модулями и оптимизирован для работы на мобильных устройствах за счет уменьшенного количества параметров.

2.8 Заключение по второй главе

Современные методы анализа, рассмотренные в главе, позволяют с разной степенью точности обнаруживать области лица и тела на кадрах, что позволяет анализировать кинематические параметры. К числу наиболее информативных показателей относятся углы суставов, траектории движения конечностей, а также производные величины — скорость и ускорение ключевых точек. Эти параметры позволяют не только анализировать текущее положение тела в пространстве, но и выявлять динамические закономерности.

Методы компьютерной обработки требуют низких вычислительных затрат, при этом имеют уступают нейронным сетям по показателям точности распознавания. В то же время нейросетевые методы не требуют ручной настройки для извлечения признаков и имеют преимущество в распознавании более сложных образов, но требуют значительно больше вычислительных ресурсов. Одним из эффективных способов является использование гибридных методов, совмещающих преимущества обоих подходов.

Таким образом, современные технологии анализа движений предоставляют ряд эффективных инструментов для объективной оценки эмоций, но требуют тщательного подбора методик в зависимости от конкретной задачи и условий съемки. Комбинация кинематических подходов открывает новые возможности для создания более точных и адаптивных систем анализа и определения эмоционального состояния.

ГЛАВА 3
ПРОЕКТИРОВАНИЕ И РЕАЛИЗАЦИЯ СИСТЕМЫ РАСПОЗНАВАНИЯ ЭМОЦИЙ

3.1 Описание приложения

Приложение представляет собой систему, которая предоставляет возможность загрузки видео и анализа видеопотока в реальном времени с камеры, определяет ключевые точки лица и тела человека, а затем классифицирует его эмоциональное состояние на основе комбинированного анализа мимики и позы. Система распознает 6 базовых эмоций: радость, грусть, удивление, гнев, отвращение и страх, а также нейтральное состояние.

3.2 Используемые технологии

Для реализации приложения был выбран язык Python, так как он предоставляет большое количество инструментов для обработки изображений и видео. Для обработки видеопотока с целью обнаружения ключевых точек лица и тела в результате сравнительного анализа был выбран фреймворк MediaPipe. Одним из преимуществ MediaPipe по сравнению с аналогами является предоставление единого API для одновременного анализа мимики лица и позы человека, что позволяет проводить комплексную оценку эмоций. Кроме того, преимуществом является возможность трекинга в реальном времени.

Также для анализа входных данных и классификации эмоций были использованы следующие библиотеки:

- NumPy — библиотека, поддерживающая работу с многомерными массивами и высокоуровневые математические функции, предназначенные для работы с многомерными массивами;
- OpenCV — библиотека алгоритмов компьютерного зрения, обработки изображений и численных алгоритмов общего назначения;
- MediaPipe — библиотека для детекции лица и позы;
- Tensorflow — библиотека для машинного обучения, используемая для решения задач построения и тренировки нейронной сети с целью классификации образов;
- Matplotlib — библиотека для визуализации данных двумерной и трхмерной графикой.

3.3 Этапы обработки видеопотока

Алгоритм работы приложения можно разделить на следующие этапы:

1. Загрузка видео / захват видео с веб-камеры;
2. Предварительная обработка кадров;
3. Детекция ключевых точек;
4. Извлечение признаков;
5. Классификация эмоций на основе извлеченных признаков.

Рассмотрим каждый из них подробнее.

3.3.1 Получение видеопотока

Приложение поддерживает два основных режима работы с видеопотоком: загрузка предварительно записанных виде в форматах MP4, MOV и захват видео в реальном времени с камеры устройства. Для этого используется библиотека OpenCV, которая последовательно получает каждый кадр и передает его на следующий этап для дальнейшей обработки.

3.3.2 Предварительная обработка кадров

Для корректной работы моделей MediaPipe используется несколько этапов предварительной обработки кадров.

Стоит отметить, что библиотека OpenCV захватывает кадры в BGR-порядке каналов, в то время как MediaPipe использует входные данные в формате RGB, что приводит к необходимости конвертации цветового пространства. В рамках реализации данного приложения нет необходимости в конвертации цветного изображения в полутоновое, так как цвет обеспечивает глубину анализа, повышая устойчивость к изменениям освещения, теням и позволяя использовать дополнительные цветовые маркеры при определении эмоций.

Еще одним преимуществом MediaPipe является отсутствие необходимости в нормализации размеров, так как это происходит автоматически, непосредственно перед обработкой кадра.

3.3.3 Детекция ключевых точек

Определение расположения ключевых точек реализуется с помощью модулей Face Mesh и Pose Estimation, предоставляемых MediaPipe (рисунок 3.1).

```
157     face_mesh = mp_face_mesh.FaceMesh(
158         max_num_faces=1,
159         refine_landmarks=True,
160         min_detection_confidence=0.5,
161         min_tracking_confidence=0.5)
162
163     pose = mp_pose.Pose(
164         min_detection_confidence=0.5,
165         min_tracking_confidence=0.5)
166
```

Рисунок 3.1 – Участок кода программы, реализующий подключение к модулям Face Mesh и Pose Estimation

В параметрах модуля распознавания лица настраиваем ограничение на детекцию 1 лица в кадре; параметр *refine_landmarks* определяет улучшенную детализацию ключевых точек (вместо 468 стандартных точек используется 478, что добавляет детализацию зрачков, важную для дополнительной информации при распознавании удивления и страха). Порог уверенности для первичной детекции позы и порог для продолжения трекинга позы определяются параметрами *min_detection_confidence* и *min_tracking_confidence* соответственно. При диапазоне значений от 0 до 1 значения установлены в 0.5 для баланса между стабильностью и чувствительностью.

3.3.4 Извлечение признаков

Этап извлечения признаков является ключевым в данном приложении, где на основе ключевых точек лица и тела вычисляются параметры, характеризующие эмоциональное состояние. Программа анализирует 478 ключевых точек лица, предоставляемых FaceMesh от MediaPipe и вычисляет параметры, представленные в таблице 3.1. Для анализа положения частей лица относительно друг друга используются точки, соответствующие уголкам глаз (133, 362), уголкам и середине губ (61, 291, 13, 14), бровям (65, 105, 295, 334), кончику носа (1, 45, 220, 295) и переносице как самой стабильной точке (6). Полный код функции, реализующей логику формирования параметров, приведет в приложении А. На рисунке 3.2 приведена развертка маски, построенной по всем ключевым точкам, с выделенными точками, используемыми для расчета параметров, с целью упрощения понимания расчетов.

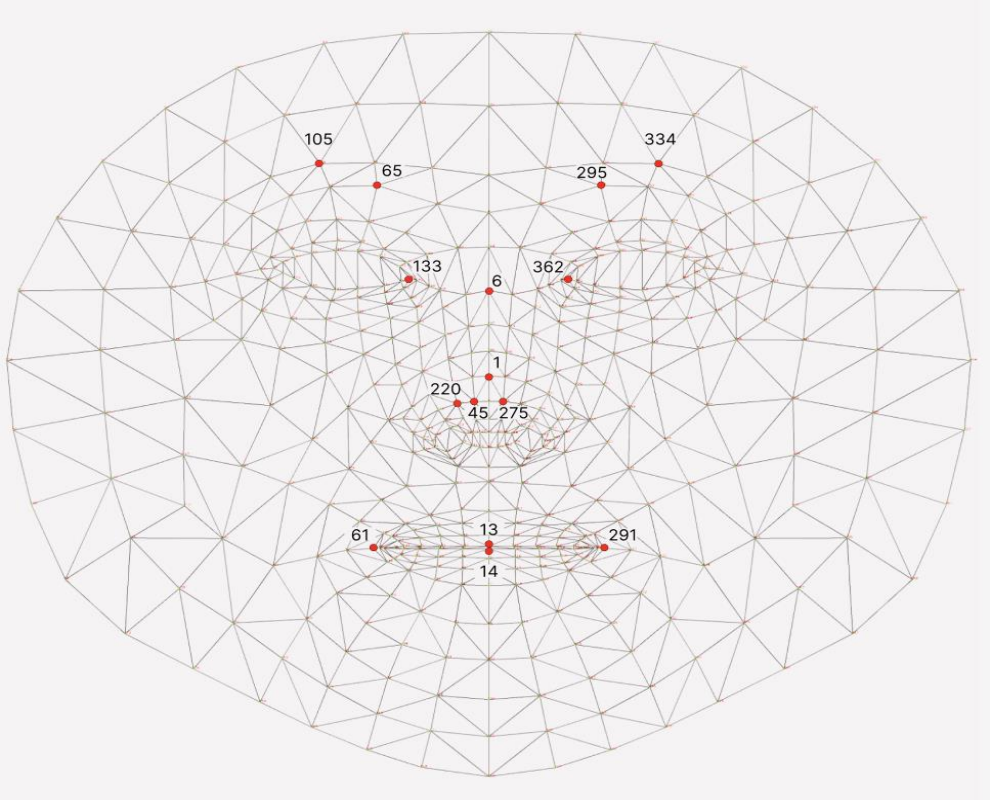


Рисунок 3.2 – Маска, формируемая ключевыми точками лица Face Mesh

Таблица 3.1 – Формулы, применяемые для вычисления параметров лица по ключевым точкам

Параметр	**Расчет**
Коэффициент улыбки	$smileCoeff = \frac{dist(points[61], points[291])}{dist(points[133], points[362])}$
Открытие рта	$mouthOpen = \frac{dist(points[13], points[14])}{dist(points[61], points[291])}$
Поднятие бровей	$browRaise = \frac{(points[105].y + points[334].y)}{points[1].y}$
Опущение уголков губ	$lipDrop = \frac{(points[105].y + points[334].y)}{points[1].y}$
Сужение бровей	$lipDrop = \frac{(points[105].y + points[334].y)}{points[1].y}$
Сморщивание носа	$mouthOpen = \frac{dist(points[13], points[14])}{dist(points[61], points[291])}$

Для определения скелета человека используется 32 ключевые точки. Более детально их взаимное расположение можно изучить на рисунке 3.3.

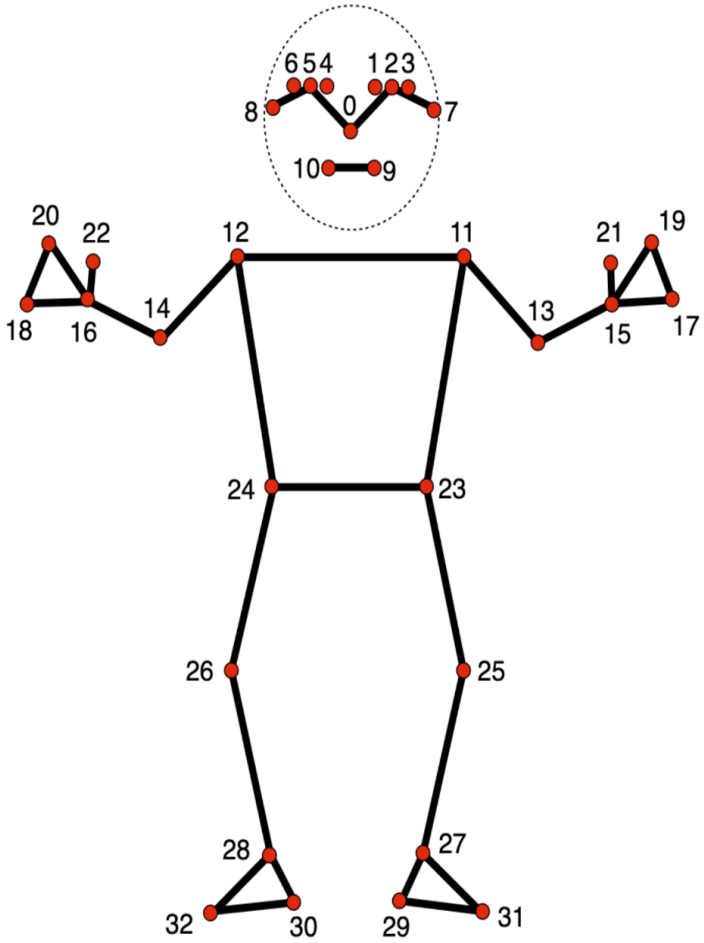


Рисунок 3.3 – Скелет, формируемый ключевыми точками тела Pose Estimation

При анализе ключевых точек тела учитывалась поза. Для эмоций радости, гнева, нейтрального состояния характерна открытая поза. Вывод о том, что поза является открытой, делается на основе вычисления угла между плечами (11, 12) и бедрами (23, 24). В то же время эмоции страха, отвращения удивления, чаще всего сопровождаются позой, в которой руки находятся перед человеком. Называемая защитной, такая поза рассчитывается на основе вычисления расстояний между левым плечом (11) и левой ладонью (15), аналогично правым плечом (12) и правой ладонью (16), после чего вычисляется среднее значение между ними.

Все вычисляемые признаки основываются на расстояниях и углах между точками. Использование отношений расстояний делает анализ устойчивым к расстоянию от человека до камеры и углу поворота головы и тела.

3.3.5 Классификация эмоций на основе извлеченных признаков

Классификатор использует значения параметров, полученные на предыдущем этапе. Классы соответствуют категориям дискретной модели эмоций, предложенной Экманом:

— радость,
— удивление,
— гнев,
— отвращение,
— печаль
— страх.

Также добавляется класс, соответствующий нейтральному состоянию. Таким образом, на выходе эмоция определяется принадлежностью к одному из 7 упомянутых выше классов.

Для решения задачи классификации было использовано два подхода:

1. Эвристический, основанный на подборе оптимальных сочетаний коэффициентов весов и порогов вручную;
2. Использование нейронной сети LSTM.

Алгоритм, используемый при первом подходе, описывается следующим образом. Сначала все параметры нормализуются, получая значение из диапазона [0,1] с использованием формулы (3.1):

$$f_{norm} = \frac{f - f_{min}}{f_{max} - f_{min}}, \tag{3.1}$$

где f – признак, извлеченный на предыдущем этапе.

Для каждой эмоции вычисляется оценка по формуле (3.2):

$$S_e = \sum_i w_i \cdot f_i, \tag{3.2}$$

где w_i — вес признака,
f_i — нормализованное значение признака.

Для интерпретации признаков используются следующие наборы: радость представляется комбинацией широкой улыбки, раскрытых зрачков и открытой позы; грусть — опущенные уголки губ, закрытая поза; удивление — поднятые брови, открытый рот; гнев — сведенные брови, сжатые губы, открытая поза; отвращение — сморщенный нос, асимметрия рта, закрытая поза; страх — поднятые брови, защитная позиция.

После подсчета оценок для каждой эмоции применяется правило максимума, то есть эмоция с максимальным значением оценки выбирается как результат:

$$emotion = arg\max_{e \in E}(S_e). \quad (3.3)$$

Далее применяется пороговая фильтрация с коэффициентом $k = 0.3$. Если значение максимальной оценки меньше указанного коэффициента, то состояние определяется как нейтральное.

Опираясь на исследования, а также различные варианты перебора соотношений коэффициентов для учета признаков лица и положения тела, наиболее точный результат показала пара значений (0.6, 0.4), где первый коэффициент определяет вклад в результирующее значение мимики, второй— позы. Более подробно зависимость точности результатов от различных сочетаний коэффициентов описана в разделах 3.6, 3.7.

Для учета временных зависимостей между кадрами классификация также проводилась с использованием LSTM. На входном слое принимается 478x3 признаков, далее сеть содержит первый LSTM-слой, возвращая последовательность для каждого временного шага и вычисляя скрытые состояния. Для предотвращения переобучения сеть содержит слой Dropout, после которого второй LSTM-слой. Дополнительное преобразование признаков происходит на следующем полносвязном слое (ReLU), далее выходной SoftMax-слой и возвращает распределение вероятностей по 7 классам эмоций.

```
model = Sequential([
    Input(shape=(None, 478, 3)),
    LSTM(128, activation='relu', return_sequences=True),
    Dropout(0.3),
    LSTM(64, activation='relu'),
    Dropout(0.3),
    Dense(32, activation='relu'),
    Dense(self.num_classes, activation='softmax')
])

optimizer = Adam(learning_rate=1e-4)
model.compile(
    optimizer=optimizer,
    loss=CategoricalCrossentropy(),
    metrics=['accuracy']
)
```

Рисунок 3.4 – Создание архитектуры LSTM-модели

Для обучения модели применялся оптимизатор Adam и функция потерь *categorical_crossentropy*, что обеспечивает эффективное обновление весов при работе с классификацией. В качестве минимального порога для определения переходных признаков используется значение 0.5.

3.4 Датасеты для обучения и тестирования

Для тестирования работы приложения использовались как видео, которые находятся в открытом доступе, так и видеопотоки в реальном времени, полученные с камеры при пользовании приложением.

AffectNet — один из крупнейших публичных датасетов, содержащий более миллиона изображений для анализа эмоциональных состояний. Включает базовые классы эмоций, предложенные Экманом. Имеет дополнительные метки валентности, отражающей эмоциональный окрас, и активации, отражающей степень выражения эмоции.

Body Language Dataset (BoLD) — набор данных с 9876 видеоклипами движений тела, является специализированным датасетом для анализа невербального поведения через распознавание поз и мимических выражений.

Тестирование с использованием камеры для распознавания эмоций в реальном времени было выполнено двумя способами: выражением эмоций с привычной изменением мимики и позы, а также путем воспроизведения характеристик каждой из эмоций согласно Системе кодирования лицевых движений. Стоит отметить, что использование второго способа дало более точные результаты определения эмоций при попытке их воспроизведения.

3.5 Оптимизация производительности

Для уменьшения нагрузки и обеспечения более плавной работы приложения в режиме реального времени были реализованы такие методы оптимизации как снижение разрешения входного видео перед обработкой. Это позволило обеспечить быструю скорость детекции с минимальной потерей точности.

Также в программе используется сохранение результатов для статичных сцен с целью уменьшения нагрузки. Таким образом, если движение в кадре не обнаружено, то на выдается предыдущий результат, что значительно снижает нагрузку на CPU.

3.6 Метрики оценки точности

Метрики являются способом количественной оценки точности. Основными задачами при разработке приложения стояла необходимость подбора оптимального сочетания коэффициентов, определяющих вклад значения эмоции, определенной по лицевым движениям, и эмоции,

выведенной из положения тела, в рамках эвристического подхода. А также оценка эффективности обучения сети.

Accuracy определяет долю правильных ответов и вычисляется по формуле (3.4):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (3.4)$$

где TP – истинно положительные предсказания (True Positive),
TN – истинно отрицательные (True Negative),
FP – ложно положительные (False Positive),
FN – ложно отрицательные (False Negative).

Для тестирования алгоритма, спроектированного самостоятельно на основе вычисления параметров по ключевым признакам, было отобрано 84 коротких видео: 12 для каждой из 7 эмоций. Каждое видео содержало отображение одной конкретной эмоции со сменой выразительности. На рисунке 3.5 представлен график, в котором отражены полученные значения точности для каждой из эмоций в результате.

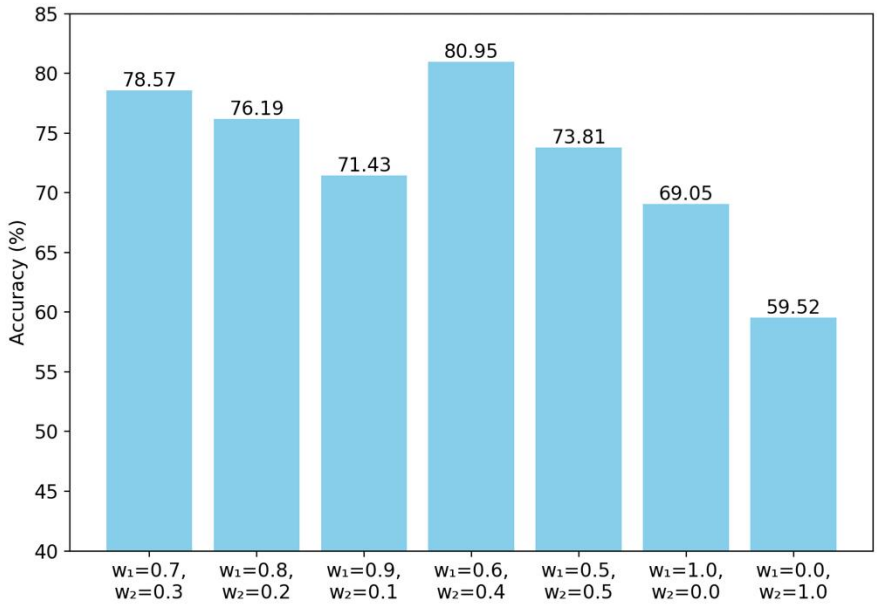


Рисунок 3.5 – Значения точности для разных сочетаний весовых коэффициентов

В рамках анализа данной системы по каждому из классов также использовались следующие метрики.

Average precision (точность) определяет долю объектов, названных положительными в рамках данного класса, ко всем положительным объектам класса и находится по формуле (3.5):

$$precision = \frac{TP}{TP + FP}. \quad (3.5)$$

Average recall (полнота) определяет долю положительных объектов, которая была обнаружена и расчитывается по формуле (3.6):

$$recall = \frac{TP}{TP + FN}. \quad (3.6)$$

F1-score – это метрика, которая объединяет значения точности и полноты в один показатель согласно формуле (3.7), помогающий оценить качество модели, что особенно актуально в случае несбалансированности данных.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}. \quad (3.7)$$

Результаты классификации при использовании сети долгой краткосрочной памяти представлены в таблице 3.

Таблица 3 – Результаты классификации LSTM

Эмоция	**Precision**	**Recall**	**F1-score**	**Кол-во примеров**
Радость	0.91	0.88	0.89	180
Удивление	0.88	0.86	0.87	130
Гнев	0.87	0.82	0.84	150
Отвращение	0.85	0.78	0.81	120
Печаль	0.81	0.83	0.82	140
Страх	0.76	0.71	0.73	100
Нейтрально	0.91	0.91	0.91	200

Учитывая несбалансированность тестовой выборки, среднее значение точности, взвешенное по количеству примеров в каждом классе, составляет 86,5%. Эмоция радости и нейтральное состояние распознаются лучше остальных, что довольно характерно для задачи распознавания эмоций. Страх часто определялся как удивление, и наоборот, что можно обусловить схожими выражениями лица при отображении этих эмоций. Эмоция грусти в силу своей слабой выраженности в ряде случаев воспринималась как нейтральное состояние.

3.7 Анализ результатов

Результаты проведенного анализа подтверждают, что положение тела человека, как в статике, так и в динамике, является одним из индикаторов, отражающих эмоциональное состояние. Выявлена следующая закономерность: точность классификации повышается при увеличении выражения эмоций.

Наиболее часто распознаваемой оказалась эмоция радости, что объясняется активным выражением этой эмоции через открытость позы. Часто эта эмоция также характеризуется поднятием рук вверх. Пример детекции точек тела и лица с определением радости на одном из кадров приведен на рисунке 3.6.

Рисунок 3.6 – Пример распознавания эмоции радости по ключевым точкам тела и лица

При определении гнева и отвращения поза имела наименьшее значение, точность распознавания только по мимическим признакам было меньше лишь на 1–2%, чем использование комбинированного анализа мимики и тела. Но общие результаты анализа учета весовых коэффициентов показывают, что анализ ключевых точек тела повышает точность распознавания эмоций. Это может быть особенно полезно в случаях, когда часть лица перекрыта аксессуарами или находится в тени.

Средняя точность, получаемая при оптимальном подборе коэффициентов весов в реализованном классификаторе, составляет 80,95%, при выборе сети долгой краткосрочной памяти – 86,5%. Стоит отметить, что поворот тела и головы вызывал ошибки в распознавании. Таким образом, можно заключить, что система эффективно распознает явные эмоции, но может быть доработана для распознавания разницы между схожими выражениями разных эмоций.

ЗАКЛЮЧЕНИЕ

В ходе дипломной работы была разработана системы для решения задачи распознавания эмоции по движениям человека в видеопотоке. Для этого последовательно был решен комплекс взаимосвязанных задач:

- Проведен анализ предметной области, что позволило сформулировать теоретическую основу для разработки алгоритм;
- Систематизированы современные методы и подходы к распознаванию эмоций, более подробно рассмотрены безмаркерные подходы в силу большей доступности в рамках проведения работы;
- Изучены и адаптированы алгоритмы извлечения ключевых точек из видеопотока;
- Выполнен сравнительный анализ инструментов для определения позы человека, на основании которого была выбраны оптимальные технологии.
- Спроектирован и реализован алгоритм работы системы, включающий этапы предварительной обработки входных данных, детекции и трекинга ключевых точек лица и тела, извлечение и классификацию эмоциональных признаков;
- Проведен сравнительный анализ точности системы, использующей эвристический подход, основанный на подсчете кинематических параметров на основе ключевых точек лица и тела, и системы, использующей сеть долгой краткосрочной памяти для классификации положения и движения тела в видеопотоке;
- Проведено тестирование системы, в результате чего были определены наиболее оптимальные параметры, используемые для вычисления признаков.

Особенностью разработанного решения стало использование комбинированных данных о движениях тела и изменениях в мимике, что позволило создать систему, устойчивую к изменениям освещения и частичным перекрытиям. MediaPipe продемонстрировал высокую эффективность в качестве фреймворка, обеспечивающего как детекцию ключевых точек, так и последующую обработку полученных данных.

Проведенные тесты показали, что система, использующая в качестве классификатора систему расчетов расстояния ключевых точек, их положение относительно друг друга достигает точности 80,95% в распознавании семи эмоциональных состояний эмоций (радость, грусть, гнев, удивление, страх, отвращение, нейтральное состояние) при работе в реальном времени. Такая система значительно уступает той, что используется LSTM для решения

задачи классификации, показывающей значения точности предсказаний 86,5%.

Разработанное решение подтверждает эффективность MediaPipe как универсального инструмента для задач анализа движений и демонстрирует перспективность подхода, основанного на кинематических характеристиках для распознавания эмоциональных состояний. Полученные результаты открывают возможности для практического применения системы в различных областях. Комбинированный подход, включающий кроме анализа мимики, анализ тела, имеет актуальность как в интерактивных развлекательных системах и маркетинговых исследованиях, так и в медицине, в рамках исследований или для диагностики заболеваний, имеющих характерные проявления в мимике, жестах и движениях.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Вайнштейн, Л. А. Общая психология / Л. А. Вайнштейн, В. А. Поликарпов, И. А. Фурманов. – Минск : Соврем. шк., 2009. – 512 с.
2. Гонсалес, Р. Цифровая обработка изображений / Р. Гонсалес, Р. Вудс. – М. : Техносфера, 2005. – 1072 с.
3. Киселев, Ю. В. Анализ подходов, методов и решений для детектирования позы человека. Выбор инструмента для задачи определения эмоционального состояния человека по его позе / Ю. В. Киселев, И. А. Богомолов, В. Л. Розалиев, В. А. Баклан // Современные наукоемкие технологии. – 2023. – № 6. – С. 41–47. – URL: https://top-technologies.ru/ru/article/view?id=39629 (дата обращения: 03.04.2025).
4. Affectiva Emotion AI: Understanding Human Emotions via Facial and Vocal Expressions. – 2023. – URL: https://www.affectiva.com/emotion-ai/ (date of access: 12.04.2025).
5. ARBEE: Towards Automated Recognition of Bodily Expression of Emotion In the Wild / Y. Luo et al. // International Journal of Computer Vision. – 2020. – Vol. 128, № 1. – P. 1–25.
6. Atkinson, A. P. Emotion perception from dynamic and static body expressions in point-light and full-light displays / A. P. Atkinson et al. // Perception. – 2004. – Vol. 33, № 6. – P. 717–746.
7. Azure Cognitive Services. – URL: https://azure.microsoft.com/ (date of access: 12.04.2025).
8. Burgoon, J. K. Nonverbal communication / J. K. Burgoon, L. K. Guerrero, K. Floyd // Handbook of communication science. – 2018. – P. 51–88. – DOI: 10.1515/9781501507920-003.
9. CMU-Perceptual-Computing-Lab OpenPose: Real-time multi-person keypoint detection library. – 2023. – URL: https://github.com/CMU-Perceptual-Computing-Lab/openpose (date of access: 02.05.2025).
10. Dalal, N. Histograms of oriented gradients for human detection / N. Dalal, B. Triggs // IEEE CVPR. – 2005. – Vol. 1. – P. 886–893.
11. Darwin, C. The expression of the emotions in man and animals / C. Darwin. – London : John Murray, 1872.
12. de Gelder, B. Why bodies? Twelve reasons for including bodily expressions in affective neuroscience / B. de Gelder // Philosophical Transactions of the Royal Society B. – 2009. – Vol. 364, № 1535. – P. 3475–3484. – DOI: 10.1098/rstb.2009.0190.

13. Ekman, P. Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life / P. Ekman. – N. Y. : Times Books, 2003. – 288 p.
14. Google MediaPipe. – URL: https://ai.google.dev/edge/mediapipe/solutions/tasks?hl=ru (date of access: 04.05.2025).
15. Jung, A. Machine Learning: The Basics / A. Jung. – Singapore : Springer, 2022. – 287 p.
16. Li, Y. Deep Learning of Human Emotion Recognition in Videos / Y. Li. – Uppsala University, 2018. – 58 p. – URL: https://www.diva-portal.org/smash/get/diva2:1174434/FULLTEXT01.pdf (date of access: 04.05.2025).
17. Mehrabian, A. Inference of attitudes from nonverbal communication in two channels / A. Mehrabian, S. R. Ferris // Journal of Consulting Psychology. – 1967. – Vol. 31, № 3. – P. 248–252.
18. Raghu, M. Transfusion: Understanding transfer learning for imaging / M. Raghu et al. // Advances in Neural Information Processing Systems. – 2019. – P. 3850.
19. Russell, J. A. A circumplex model of affect / J. A. Russell // Journal of Personality and Social Psychology. – 1980. – Vol. 39, № 6. – P. 1161–1178.
20. Schlosberg, H. Three dimensions of emotion / H. Schlosberg // Psychological Review. – 1954. – Vol. 61, № 2. – P. 81–88.
21. Shanmugamani, R. Deep Learning for Computer Vision / R. Shanmugamani. – Packt Publishing, 2018. – 304 p.
22. Sun, X. Classification for Remote Sensing Data with Improved CNN Method / X. Sun et al. // IEEE Access. – 2019. – URL: https://www.researchgate.net/publication/337177628_Classification_for_Remote_Sensing_Data_with_Improved_CNN_Method (date of access: 15.12.2024).
23. Wallbott, H. G. Bodily expression of emotion / H. G. Wallbott // European Journal of Social Psychology. – 1998. – Vol. 28, № 6. – P. 879–896.
24. Wei, S.-E. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields / S.-E. Wei et al. // arXiv. – 2016. – URL: https://arxiv.org/abs/1611.08050 (date of access: 02.05.2025).
25. Woods, K. Combination of Multiple Classifiers Using Local Accuracy Estimates / K. Woods, W. P. Kegelmeyer. – IEEE, 1997. – 410 p.

ПРИЛОЖЕНИЕ А

КОД МОДУЛЯ ПРОГРАММЫ, РЕАЛИЗУЮЩЕГО ФОРМИРОВАНИЕ ПАРАМЕТРОВ В РАМКАХ ЭВРИСТИЧЕСКОГО ПОДХОДА

```
def process_frame(image, face_mesh, pose):
    image_rgb = cv2.cvtColor(image, cv2.COLOR_BGR2RGB)
    results_face = face_mesh.process(image_rgb)
    results_pose = pose.process(image_rgb)

    h, w = image.shape[:2]
    emotion = "neutral"

    if results_face.multi_face_landmarks:
        face_landmarks = results_face.multi_face_landmarks[0]
        landmarks = face_landmarks.landmark
        points = [(int(lm.x * w), int(lm.y * h)) for lm in landmarks]

        mouth_width = math.dist(points[61], points[291])
        eye_distance = math.dist(points[133], points[362])
        smile_coeff = mouth_width / eye_distance if eye_distance > 0 else 0
        param_history['smile_coeff'].append(smile_coeff)

        mouth_height = math.dist(points[13], points[14])
        mouth_open_ratio = mouth_height / mouth_width if mouth_width > 0 else 0
        param_history['mouth_open'].append(mouth_open_ratio)

        brow_raise = (points[105][1] + points[334][1])/2 - points[1][1]
        param_history['brow_raise'].append(brow_raise)

        lip_drop = points[14][1] - (points[61][1] + points[291][1])/2
        param_history['lip_drop'].append(lip_drop)

        brow_frown = math.dist(points[65], points[295])
        param_history['brow_frown'].append(brow_frown)

        base_distance = math.dist(points[45], points[275])
        nose_wrinkle = math.dist(points[6], points[220]) / base_distance if base_distance > 0 else 0
        param_history['nose_wrinkle'].append(nose_wrinkle)

    if results_pose.pose_landmarks:
        pose_landmarks = results_pose.pose_landmarks
        pose_points = [(int(lm.x * w), int(lm.y * h)) for lm in pose_landmarks.landmark]

        shoulders_vec = np.array(pose_points[12]) - np.array(pose_points[11])
        hips_vec = np.array(pose_points[24]) - np.array(pose_points[23])
        shoulder_hip_angle = np.degrees(np.arccos(
            np.dot(shoulders_vec, hips_vec) / (np.linalg.norm(shoulders_vec) * np.linalg.norm(hips_vec))
        )) if np.linalg.norm(shoulders_vec) * np.linalg.norm(hips_vec) > 0 else 0
        param_history['pose_openness'].append(shoulder_hip_angle / 180)

        left_defense = math.dist(pose_points[15], pose_points[11])
        right_defense = math.dist(pose_points[16], pose_points[12])
        defense_pos = (left_defense + right_defense) / 2
        param_history['defense_pos'].append(defense_pos / 100)
```