

Система обнаружения девиантного поведения по видеопоследовательностям лица человека на транспорте

С.А. Онопко, Ю.А. Павелина, Ю.И. Минкин, П.А. Попов

Аннотация — В работе рассматривается задача обнаружения девиантного поведения по видеопоследовательностям лица человека. В отличие от статического анализа отдельных кадров, исследование опирается на временные паттерны мимики и физиологические состояния, наблюдаемые на протяжении всего видео. Для проверки гипотезы о применимости визуально-языковых моделей к интерпретации видео использовалась модель gpt5.5-quick, а результаты оценки фиксировались по двум группам признаков: эмоции и состояния. Показано, что система обладает достаточной точностью работы для решения прикладных задач автоматизированного анализа эмоционального состояния и обнаружения признаков отклоняющегося поведения по видеоданным. Результаты подтверждают работоспособность подхода на уровне проверки концепции; на следующем этапе предполагается локальное развертывание открытой визуально-языковой модели на собственном, более производительном вычислительном оборудовании.

Ключевые слова — девиантное поведение, техническое зрение, нейронные сети.

I. ВВЕДЕНИЕ

Задача обнаружения девиантного поведения по паттернам лица человека относится к классу видеозадач, в которых важна не только пространственная конфигурация лица, но и динамика ее изменения. Это особенно важно для прикладных сценариев, связанных с мониторингом объектов транспортной инфраструктуры, где своевременное выявление аномального состояния человека может иметь практическую значимость.

Современные методы анализа лица в видео показывают [1, 2], что выражения эмоций и физиологических состояний проявляются как последовательности с характерными фазами: переход от базовой конфигурации, пик выражения и последующее удержание или возврат к норме [3]. Именно поэтому анализ видеопоследовательности дает больше информации, чем одиночный кадр.

II. ПОСТАНОВКА ЗАДАЧИ

В рамках разработки систем обнаружения девиантного поведения требовалось проверить способность моделей нейронных сетей корректно интерпретировать эмоции и поведенческие состояния по видео и оценить качество

предсказаний. В исследовании рассматривались два типа выходов: предсказание одной доминирующей эмоции и одного доминирующего состояния для всего видео.

Для эмоций использовались классы: «нейтральное состояние», «злость», «отвращение», «страх», «удивление», «счастье».

Для состояний использовались классы: «нейтральный», «глаза закрыты», «сонный/зевок», «блуждающий взгляд», «комбинация аномалий». Такой набор классов позволяет описывать как базовые эмоциональные паттерны, так и физиологически значимые аномалии лица.

III. МЕТОДИКА АНАЛИЗА

В основе анализа лежал принцип интерпретации всей последовательности видео, а не одного кадра. Для эмоций ключевыми были устойчивость выражения, наличие перехода от базовой мимики к экспрессивной, а также доминирующая зона активации лица. Для состояний важны были длительность, повторяемость и сочетание признаков, например длительное закрытие глаз или сочетание асимметрии лица с низкой активностью взгляда.

Для проверки подхода в окончательном эксперименте использовалась модель gpt5.5 [4] как наиболее результативная из опробованных. Предварительные эксперименты с другими VLM (Visual Language Model) показали, что альтернативные модели не давали улучшения качества по сравнению с ней.

IV. ЭКСПЕРИМЕНТАЛЬНАЯ МЕТОДОЛОГИЯ И ДАТАСЕТ

Для оценки качества моделей был сформирован комбинированный датасет, включающий как собственные видеозаписи, так и данные из открытых источников. Собственная часть датасета содержит по 17 видео на каждый из 6 классов эмоций и по 17 видео на каждый из 4 классов физиологических состояний: всего 170 видеозаписей, снятых на веб-камеру с частотой не менее 25 кадров в секунду в разрешении не ниже 640×480. Все собственные видео постановочные. Съемка велась при равномерном освещении без сильных теней, лицо занимало 15–30% площади кадра, положение головы фронтальное с допустимыми небольшими поворотами. Средняя длительность видеозаписи для эмоций составила 7,84 секунды, для состояний от 5–10 секунд («закрытые глаза», «блуждающий взгляд», «комбинация аномалий») до 30–60 секунд («сонный/зевок»), поскольку для устойчивого выявления

сонливости требуется наблюдение не менее двух эпизодов зевания в течение 60 секунд.

Дополнительно привлекались фрагменты из открытых датасетов: FAMED, KMU-FED и MLI-DER для класса эмоций, и YawDD для класса «сонный/зевок». Использование открытых датасетов было ограничено: азиатские датасеты (KMU-FED, MLI-DER) не использовались в финальном тесте из-за отличий в мимических паттернах от целевой группы наблюдения, а датасет DFEW был отклонен, поскольку представленные в нем ракурсы съемки (кинематографические кадры) не соответствуют условиям наблюдения на транспортных объектах (фронтальная камера видеонаблюдения).

Итоговая тестовая выборка, на которой рассчитаны метрики, приведенные в разделе «Эмпирическое исследование», состоит из 170 собственных видеозаписей, сбалансированных по классам (по 17 видео на класс), что исключает смещение метрик в сторону доминирующего класса. Отдельная обучающая выборка не формировалась, так как в работе используется VLM в режиме вывода без дообучения. Соответственно, все 170 видео использовались как тестовая (эталонная) выборка, а валидация промпта проводилась итеративно на отдельном пилотном подмножестве из 20 видео, не входящем в финальную тестовую выборку.

V. ТЕХНИЧЕСКИЕ ДЕТАЛИ ПРИМЕНЕНИЯ VLM

Модель gpt5.5-quick применялась в режиме few-shot вывода без дообучения весов (zero-shot дообучение не проводилось; в промпт включались краткие текстовые описания признаков каждого класса, аналогичные приведенным в Таблицах 1 и 2, что соответствует постановке few-shot). Формат промпта включал:

1. системную инструкцию с перечислением целевых классов эмоций и состояний и их визуальных признаков;
2. явное указание приоритета выраженной кратковременной эмоции над более длительным нейтральным состоянием. Данное указание было введено по результатам пилотного тестирования, показавшего склонность модели «сглаживать» кратковременные экспрессивные сигналы;

3. требование вернуть структурированный ответ в формате JSON, содержащий вектор вероятностей по всем классам целевой группы (эмоции или состояния) для видео в целом, без поккадровой разметки.

Модель обрабатывает видео целиком (не поккадрово) и выдает вероятностное распределение по классам, а не свободный текстовый ответ, что позволяет напрямую сопоставлять ее выход с эталонной разметкой и рассчитывать Precision, Recall и F1-score стандартным образом. Такая постановка не привязана к конкретной модели: промпт и формат ответа могут использоваться с другими VLM, способными обрабатывать видео и возвращать структурированный ответ.

Выбор модели gpt5.5-quick обусловлен результатами сравнительного тестирования на собранном датасете (170 видео) среди нескольких VLM (Visual Language Model), включавших Qwen3-VL, Molmo2, InternVL3.5-8B и Gemma. По итогам сравнения на промежуточном этапе лучшие результаты показала модель Qwen-VL (по совокупности классов эмоций и состояний: Qwen VL >

Molmo > InternVL > Gemma), однако при последующем расширении экспериментов и переходе на gpt5.5-quick данная модель показала более высокую устойчивость на полном наборе классов и была использована в финальном эксперименте по проверке концепции (proof of concept), результаты которого приведены в разделе «Эмпирическое исследование».

Попытки использования более крупных открытых моделей (Qwen3-VL-8B-Instruct-AWQ-4bit, InternVL3.5-8B) в доступной вычислительной инфраструктуре не привели к улучшению метрик и в ряде случаев были ограничены аппаратными ресурсами. Поэтому на следующем этапе исследования планируется локальное развертывание открытой VLM (например, из семейства Qwen-VL) на собственном, более производительном вычислительном оборудовании, на котором можно запускать модели большего размера. Это позволит выполнять интерпретацию видео без обращения к внешним облачным сервисам, что необходимо для тиражирования решения в реальных условиях эксплуатации.

Специализированные архитектуры для анализа временных видеопоследовательностей (EfficientNet совместно с LSTM, TimeSformer, VideoMAE) не были включены в сравнение на данном этапе исследования по следующим причинам: во-первых, применение VLM в режиме few-shot вывода не требует обучающей выборки и дообучения, что существенно снижает трудоемкость и время внедрения прикладного решения на данных небольшого объема (170 видео недостаточно для устойчивого обучения специализированной архитектуры с нуля); во-вторых, задачей текущего этапа являлась оценка принципиальной применимости VLM-подхода для интерпретации видеопоследовательностей лица в задаче обнаружения девиантного поведения. Сравнение с дообучаемыми специализированными архитектурами (EfficientNet совместно с LSTM, TimeSformer, VideoMAE), обученными на расширенном датасете, запланировано как направление дальнейшего исследования.

VI. КЛАССИФИКАЦИЯ ЭМОЦИЙ ЧЕЛОВЕКА И ЕГО СОСТОЯНИЙ

A. Эмоции лица

Для анализа эмоций наиболее устойчивыми оказались классы «нейтральный» и «счастье». Это объясняется тем, что нейтральное состояние имеет стабильную базовую конфигурацию лица, а «счастье» проявляется через отчетливое поднятие уголков рта и активизацию щек, что делает этот класс визуально более отделимым.

При слабом или кратковременном выражении модель склонна «сглаживать» сигнал и относить эмоцию к «нейтральному состоянию». Типичная группа ошибок – ошибка классификации между «страх» и «удивление», а также между «злость» и «отвращение». Признаки обнаружения различных классов эмоций приведены в Таблице 1.

Таблица 1 – Признаки обнаружения классов эмоций

Класс эмоции	Признаки обнаружения
Нейтральное состояние	Преобладание спокойного выражения лица, отсутствие устойчивого напряжения лицевых мышц, положение бровей и рта близко к естественному состоянию.
Злость	Опущенные и сведенные к центру брови, выраженная межбровная складка, напряжение области глаз, сужение глазной щели.
Отвращение	Сморщивание носа, поднятие верхней губы, напряжение носогубной области, поднятие щек, сужение глаз.
Страх	Устойчиво широко раскрытые глаза, увеличенная видимость склеры, приподнятые брови, длительное сохранение данного выражения лица.
Удивление	Быстрое расширение глаз, поднятие бровей, увеличение раскрытия век, резкий переход от нейтрального состояния к активному выражению лица.
Счастье	Поднятые уголки рта, улыбка, приподнятые щеки, преобладание положительной мимики нижней части лица.

По итогам тестирования лучшую результативность среди опробованных моделей для определения эмоций показала модель gpt5.5, однако общая картина по матрице ошибок показывает, что классы «страх», «удивление», «злость» и «отвращение» остаются наиболее сложными с точки зрения строгого разделения.

В. Поведенческие состояния

Для состояний наиболее уверенно определялись закрытие глаз и зевание. Эти классы обладают выраженной временной структурой: длительное закрытие глаз должно сохраняться не менее нескольких секунд, а зевота проявляется как повторяющийся цикл с отчетливым открытием рта. Такие паттерны легче отделяются от нормального поведения, чем более тонкие аномалии. Признаки обнаружения различных классов состояний приведены в Таблице 2.

Таблица 2 – Признаки обнаружения классов состояний

Класс состояния	Признаки обнаружения
Нейтральное состояние	Отсутствие признаков аномальных состояний, естественная мимическая активность и движение глаз.
Закрытые глаза	Непрерывное закрытие обоих глаз на протяжении не менее 3 секунд.
Сонный/зевок	Не менее двух эпизодов широкого открытия рта и опускания нижней челюсти в течение 60 секунд.

Блуждающий взгляд	Частые изменения направления взгляда, переключение между различными точками фиксации при сохранении относительно нейтрального выражения лица.
Комбинация аномалий	Устойчивая асимметрия лица (одностороннее напряжение, асимметричная улыбка и др.) в сочетании с длительно низкой активностью движения взгляда.

Состояние «блуждающий взгляд» распознавалось хуже, так как связано не с одной яркой мимической вспышкой, а с устойчивым характером движения глаз и взгляда. Класс «комбинация аномалий» оказался самым сложным, поскольку требует одновременного наличия асимметрии лица и низкой активности взгляда в течение длительного времени.

Таким образом, состояния с простыми и повторяющимися визуальными признаками обладают более высокой различимостью, тогда как комбинированные и маловыраженные аномалии остаются проблемными для текущих моделей.

VII. ЭМПИРИЧЕСКОЕ ИССЛЕДОВАНИЕ

А. Результаты эксперимента

По результатам тестирования системы на задаче распознавания эмоций достигнуто среднее значение Precision = 0,767, Recall = 0,725 и F1-score = 0,731. Наиболее качественно распознается класс «счастье» (F1-score = 0,970), что свидетельствует о высокой различимости признаков данного эмоционального состояния. Также высокие показатели получены для класса «отвращение» (F1-score = 0,786). Наиболее сложным для распознавания оказался класс «страх» (F1-score = 0,553), что может быть связано со сходством его визуальных признаков с другими эмоциональными состояниями. В целом результаты демонстрируют удовлетворительное качество распознавания эмоциональных состояний на видеопоследовательностях.

В задаче классификации физиологических состояний достигнуты средние значения Precision = 0,855, Recall = 0,788 и F1-score = 0,778, что превышает показатели распознавания эмоций. Лучшие результаты были получены для классов «закрытие глаз» (F1-score = 0,970) и «сонный/зевок» (F1-score = 0,903), признаки которых обладают высокой визуальной выраженностью. Для класса «блуждающий взгляд» достигнута максимальная полнота обнаружения (Recall = 1,000), что свидетельствует об отсутствии пропусков объектов данного класса. Наибольшую сложность представляет класс «комбинация аномалий» (F1-score = 0,522), объединяющий несколько признаков и требующий одновременного выполнения нескольких условий обнаружения.

Сводные метрики показывают, что задача в целом

решается на уровне, пригодном для исследовательской постановки, но требует дальнейшей доработки промптов, отдельных правил для эмоций и состояний, а также улучшения обработки пограничных классов.

В. Анализ ошибок

Основная ошибка для классификации эмоций – подмена экспрессивных классов нейтральным. Это указывает на то, что модель лучше фиксирует устойчивую форму лица, чем слабую динамику выражения. Между классами «страх» и «удивление» ошибка особенно типична, поскольку оба класса связаны с расширением глаз и поднятием верхней части лица, но различаются именно по временной структуре и характеру перехода.

Для состояний типичен сдвиг в сторону «нейтральный», что отражает осторожность модели при недостаточной уверенности. Это полезно с точки зрения снижения ложноположительных ошибок, но создает риск пропуска слабовыраженных аномалий.

VIII. ЗАКЛЮЧЕНИЕ

В работе показано, что паттерны лица содержат достаточную информацию для обнаружения девиантного поведения и аномальных физиологических состояний. Обнаружение девиантного поведения по паттернам лица человека возможно при анализе видеопоследовательности и учете временной структуры мимики.

При совместной оценке всех классов эмоций и состояний достигнуты средние показатели Precision = 0,807, Recall = 0,757 и F1-score = 0,754. Полученное значение F1-score выше 0,75 свидетельствует о достаточной точности работы системы для решения прикладных задач автоматизированного анализа эмоционального состояния и обнаружения признаков отклоняющегося поведения по видеоданным.

Результаты, полученные на постановочных записях, подтверждают гипотезу о применимости VLM к интерпретации видеопоследовательностей лица на уровне лабораторной проверки концепции. Модель gpt5.5-quick использовалась как инструмент этой проверки, а достигнутые с ее помощью метрики станут ориентиром при переходе на открытую VLM, развернутую локально на собственном, более производительном вычислительном оборудовании.

БЛАГОДАРНОСТИ

Работа выполнена в РУТ (МИИТ) в рамках государственного задания № 103-00001-26-00 по проведению прикладных научных исследований по теме «Разработка методов и технологии интеллектуальной системы анализа видеоизображений с целью выявления лиц с признаками девиантного и социально опасного поведения на объектах транспортной инфраструктуры».

БИБЛИОГРАФИЯ

- [1] Jeganathan J. et al. Quantifying dynamic facial expressions under naturalistic conditions //Elife. – 2022. – Т. 11. – С. e79581.
- [2] Zhao J. et al. Facial dynamics in video: Instruction tuning for improved facial expression perception and contextual awareness //Proceedings of the AAAI Conference on Artificial Intelligence. – 2026. – Т. 40. – № 16. – С. 13163-13171.
- [3] Cuve H. C. J., Sowden-Carvalho S., Cook J. Spatiotemporal structure and substates in emotional facial expressions //bioRxiv. – 2024. – С. 2024.09. 02.610622.
- [4] Wu L. et al. RoboStressBench: Benchmarking VLM Robustness to Physical Visual Stress in Embodied Scenes //arXiv preprint arXiv:2606.00828. – 2026.

Статья получена 30 июля 2026.

Онопко София Александровна, ведущий инженер-программист отдела искусственного интеллекта Отделения систем компьютерного зрения Центра интеллектуальных транспортных систем и автономных транспортных средств НИИТ, Российский университет транспорта, email: onorkosonja@yandex.ru

Павелина Юлия Александровна, начальник сектора системного анализа Отделения реализации проектов Центра интеллектуальных транспортных систем и автономных транспортных средств НИИТ, Российский университет транспорта, email: lyakhovenko.kam@gmail.com

Минкин Юрий Игоревич, к.т.н., заведующий отделением реализации проектов Центра интеллектуальных транспортных систем и автономных транспортных средств НИИТ, Российский университет транспорта, email: yury.i.minkin@yandex.ru

Попов Павел Александрович, к.т.н., начальник Центра интеллектуальных транспортных систем и автономных транспортных средств НИИТ, Российский университет транспорта, email: popovpavela@yandex.ru

A system for detecting deviant behavior based on facial video sequences in public transport

S.A. Onopko, J.A. Pavelina, J.I. Minkin, P.A. Popov

Abstract — This paper examines the problem of detecting deviant behavior in facial video sequences. Unlike static analysis of individual frames, the study relies on temporal patterns of facial expressions and physiological states observed throughout the video. To test the hypothesis that vision-language models are applicable to video interpretation, the gpt5.5-quick model was used, and the evaluation results were recorded using two groups of features: emotions and states. It is demonstrated that the system has sufficient accuracy for solving applied problems of automated emotional state analysis and detecting signs of deviant behavior in video data. The results confirm that the approach works at the proof-of-concept level; the next stage involves local deployment of an open vision-language model on more powerful in-house hardware.

Keywords — deviant behavior, machine vision, neural networks.

REFERENCES

- [1] Jeganathan J. et al. Quantifying dynamic facial expressions under naturalistic conditions //Elife. – 2022. – T. 11. – C. e79581.
- [2] Zhao J. et al. Facial dynamics in video: Instruction tuning for improved facial expression perception and contextual awareness //Proceedings of the AAAI Conference on Artificial Intelligence. – 2026. – T. 40. – № 16. – C. 13163-13171.
- [3] Cuve H. C. J., Sowden-Carvalho S., Cook J. Spatiotemporal structure and substates in emotional facial expressions //bioRxiv. – 2024. – C. 2024.09. 02.610622.
- [4] Wu L. et al. RoboStressBench: Benchmarking VLM Robustness to Physical Visual Stress in Embodied Scenes //arXiv preprint arXiv:2606.00828. – 2026.