УДК 004.93

ИНТЕЛЛЕКТУАЛЬНЫЙ СЕРВИС МУЛЬТИМОДАЛЬНОГО НЕЙРОСЕТЕВОГО МОНИТОРИНГА ОБЛАСТИ НАБЛЮДЕНИЯ

Р. Р. Миннеахметов[0009-0007-8551-1393]

Казанский (Приволжский) федеральный университет, г. Казань, Россия
razil0071999@gmail.com

Аннотация

Представлен подход к разработке интеллектуального сервиса мультимодального мониторинга области наблюдения с использованием больших нейросетевых моделей. Предлагаемое решение способно анализировать разнородные данные: видеопотоки, сигналы датчиков окружающей среды (температура, влажность и пр.) и журналы событий – для получения целостной картины происходящего. В качестве основных инструментов задействованы крупные языковые и визуальные модели (например, LLaMA, MiniCPM-V и др.), развернутые локально с помощью платформы Ollama, что обеспечивает автономную и безопасную обработку информации без необходимости передачи данных на удаленные сервера. Разработан прототип системы, работающий в офлайн-режиме и способный выявлять критические ситуации, аномальные отклонения от нормы и контекстно значимые события в наблюдаемой зоне. Описана методика формирования тестовых сценариев и проведения качественной оценки работы модели по метрикам F1-мера, Precision, Recall. Результаты экспериментов подтвердили применимость мультимодальных моделей для решения задач мониторинга: прототип успешно распознает сложные паттерны поведения и демонстрирует потенциал больших моделей в построении адаптивных и масштабируемых систем наблюдения.

Ключевые слова*: интеллектуальный сервис, мультимодальный мониторинг, Ollama, большие языковые модели, отслеживание активностей, видеоаналитика, искусственный интеллект.*

ВВЕДЕНИЕ

В современном мире наблюдается стремительный рост объема данных, генерируемых различными сенсорами, системами видеонаблюдения и другими устройствами интернета вещей. Это стимулирует развитие интеллектуальных систем, которые все чаще применяются для анализа поведенческих и ситуационных паттернов в реальном времени. Одним из перспективных направлений в этой области является мультимодальный мониторинг – анализ информации, поступающей одновременно из различных источников (видео, датчики, логи и т. д.) [1]. Благодаря объединению разнородных данных такой подход позволяет получить более полную и достоверную картину происходящего за счет перекрестной верификации сведений из разных модальностей. Под областью наблюдения в контексте настоящей работы понимается ограниченное пространство (физическое или логическое), в котором ведется автоматизированное отслеживание активности. Это может быть помещение, коридор, производственный участок или виртуальная зона, которая контролируется с помощью видеокамер, сенсоров либо систем логирования событий.

Традиционные системы безопасности и мониторинга, как правило, основаны на сигнатурном анализе и ручной настройке правил срабатывания. Они эффективно выявляют известные угрозы, но могут не замечать новые или нетипичные ситуации. В отличие от таких подходов, интеллектуальный сервис, предлагаемый в настоящей работе, использует возможности больших нейросетевых моделей, как языковых, так и визуальных, для автономного анализа происходящих событий. Большие языковые модели (Large Language Models, LLM) и современные модели компьютерного зрения (Vision-модели) демонстрируют высокую эффективность в самых различных задачах: обработке естественного языка, распознавании образов, анализе временных рядов и т. д. Их применение в системах мониторинга позволяет автоматизировать распознавание сложных поведенческих паттернов и потенциально опасных действий, что ранее требовало значительных вычислительных ресурсов и вмешательства человека [2]. В проактивных системах кибербезопасности похожие методы уже используются для поиска аномалий, не выявляемых стандартными средствами защиты [3]. Под ак-

тивностью в общем случае понимается любое значимое изменение в наблюдаемой зоне – будь то событие, зафиксированное системой логирования, либо действие, зафиксированное на видеокамере.

Настоящая работа направлена на создание прототипа интеллектуальной системы мониторинга, способной локально (на персональном устройстве) анализировать мультимодальные данные активности и выявлять важные события. Особое внимание уделено архитектуре и идее системы, основанной на внедрении больших предобученных моделей для анализа нескольких типов данных одновременно, с акцентом на автономность и безопасность обработки. Предварительные результаты работы были представлены в виде доклада на научной конференции «Научный сервис в сети Интернет» [4]; в настоящей статье эти материалы существенно расширены и углублены. Более подробно рассмотрены структура предлагаемого решения, используемые модели и методы, экспериментальные сценарии и полученные результаты.

1. ОБЗОР СУЩЕСТВУЮЩИХ РЕШЕНИЙ

Для эффективного отслеживания активностей в сложных условиях требуется анализировать информацию из различных источников: видеопотоки, системные логи, показания датчиков, а также данные, генерируемые пользователями (User Generated Content). Современные решения предлагают использовать для этого предобученные большие нейросетевые модели, способные извлекать значимые паттерны из разнородных входных данных [5]. В области компьютерного зрения широко применяются глубокие сверточные нейронные сети и трансформеры Vision Transformer для распознавания действий на видео и классификации поведения людей в режиме реального времени [6]. Например, решения на базе CNN и Vision Transformer успешно идентифицируют различные виды человеческой активности на видеозаписях (ходьба, бег, падение и т. п.) и могут обнаруживать отклонения от нормы [7].

Для анализа последовательностей сигналов носимых и окружающих сенсоров (акселерометров, гироскопов, датчиков среды) применяют рекуррентные архитектуры (LSTM/GRU) или трансформеры, обученные на больших массивах

данных о движениях людей. Такие модели способны выявлять характерные последовательности сигналов, соответствующие определенным видам активности, и обнаруживать нетипичные паттерны, сигнализирующие о возможном инциденте [7]. В частности, в задачах распознавания человеческой активности (Human Activity Recognition, HAR) по данным носимых устройств большие нейросети достигли заметных успехов [8]. Они позволяют в реальном времени отслеживать показатели движений и состояния здоровья, генерируя уведомления при выявлении опасных событий (например, резкое падение человека, приступ аритмии и т. д.).

Для текстовых данных (таких как журналы событий, протоколы и отчеты) все шире используются большие языковые модели трансформерного типа. Они способны интерпретировать последовательности записей как связный текст и по контексту выявлять аномалии или критические события [9]. В сфере кибербезопасности подобные модели анализируют сетевые логи и сообщения, распознавая характерные предвестники атак и киберинцидентов, что повышает оперативность реагирования на угрозы [10].

Сочетание мультимодального анализа данных на базе больших моделей уже находит применение в различных прикладных областях. В промышленности системы компьютерного зрения на основе глубоких нейросетей контролируют соблюдение техники безопасности на производстве например автоматически обнаруживают отсутствие каски или спецодежды у рабочего [11]. Анализ вибраций и других сенсорных данных станков с помощью рекуррентных нейросетей позволяет реализовать предиктивное обслуживание оборудования (predictive maintenance), выявляя отклонения в работе механизмов и предотвращая аварии [12]. В сфере «умных» домов крупные модели мониторят повседневную активность жильцов для повышения комфорта и безопасности: так, по данным камер и датчиков движения можно определить, что пожилой человек упал, и автоматически вызвать помощь [13]. Носимые фитнес-устройства с интегрированными моделями HAR отслеживают физическую активность и состояние здоровья пользователя, сигнализируя при обнаружении аномалий (например, чрезмерно длительной неподвижности или аритмии) [8]. В системах общественной безопасности нейросетевые алгоритмы видеоаналитики способны распознавать по-

дозрительные действия в режиме реального времени: оставленные без присмотра предметы, агрессивное поведение в толпе, тем самым помогая предотвращать правонарушения и инциденты [10]. Еще одним направлением применения крупных моделей являются медицина и здоровье: обработка потоков данных от носимых сенсоров и даже анализ речи/текста пациентов (записи сессий, соцсети) с помощью LLM дают возможность выявлять признаки стресса, депрессии или ухудшения физического состояния на ранних стадиях [2, 13].

Таким образом, достижения последних лет демонстрируют универсальность и эффективность больших нейросетевых моделей в задачах мониторинга: от производственных цехов до домашней обстановки они позволяют повысить качество наблюдения и снизить влияние человеческого фактора. Вместе с тем многие существующие решения либо сфокусированы на одной модальности данных, либо требуют значительных ресурсов и предварительной настройки под конкретные сценарии. Актуальной задачей остается разработка единой интеллектуальной системы, способной интегрировать несколько источников данных и автоматически выявлять сложные ситуации без заранее прописанных правил. В следующем разделе формально описана постановка задачи для такого сервиса.

2. ПОСТАНОВКА ЗАДАЧИ

Цель настоящего исследования состояла в следующем: разработать прототип интеллектуальной системы мониторинга, способной локально в автономном режиме анализировать мультимодальные данные активности (видеоизображения, показания сенсоров, текстовые логи) и своевременно обнаруживать потенциально опасные или аномальные ситуации. В цель работы входила также оценка применимости крупных предобученных нейросетевых моделей для отслеживания различных видов активности в реальных сценариях и выработки соответствующей реакции на выявленные события.

Для достижения этой цели решались следующие задачи: во-первых, реализовать локальное развертывание современных больших моделей (языковых и визуальных) и обеспечить их совместную работу с различными типами входных данных; во-вторых, разработать набор тестовых сценариев, имитирующих

типичные ситуации в области наблюдения (чрезвычайные происшествия, нештатные события и штатный режим), чтобы проверить работоспособность системы; в-третьих, провести сравнительную оценку нескольких моделей по точности распознавания ситуаций и производительности (времени отклика) и на этой основе определить оптимальные решения и узкие места прототипа.

Отметим, что хотя мультимодальные системы теоретически могут включать анализ звука и речи, в рамках настоящей работы аудиомодальность не рассматривается. Это обусловлено, с одной стороны, ограниченной поддержкой аудиовходов в большинстве доступных LLM- и Vision-моделей (на момент исследования), а с другой – отсутствием звуковых данных во многих системах видеонаблюдения (звук обычно не записывается). Тем не менее заложенная архитектура сервиса допускает расширение за счет подключения дополнительных модальностей, включая звук или биометрические датчики, при наличии соответствующих моделей и аппаратуры.

3. ЛОКАЛЬНОЕ РАЗВЕРТЫВАНИЕ МОДЕЛЕЙ С ПОМОЩЬЮ OLLAMA

Для выполнения поставленной задачи было решено использовать локальное развертывание больших нейросетевых моделей, что обеспечивает автономность и конфиденциальность обработки данных. В прототипе использован инструмент Ollama – легковесная платформа, позволяющая запускать различные предобученные LLM- и Vision-модели на персональном компьютере и взаимодействовать с ними через простой интерфейс. Ollama поддерживает современные архитектуры моделей (семейства LLaMA, Mistral и др.) и предоставляет гибкий REST API для их интеграции [14]. Одним из ключевых преимуществ Ollama является возможность полностью локальной работы: все вычисления происходят на стороне пользователя, без отправки входных данных (например, видеокадров или логов) на удаленные серверы. Это особенно важно при работе с чувствительной информацией, требующей соблюдения политики безопасности и приватности [15].

Взаимодействие с моделью в Ollama осуществляется путем отправки HTTP-запросов на локальный сервер (по умолчанию – порт 11434). Запрос формируется в формате JSON и включает обязательные поля:

- **model** – идентификатор выбранной модели (название веса LLM/Vision-модели, загруженной в Ollama);
- **prompt** – текст инструкции или вопрос, передаваемый модели;
- **temperature** – параметр стохастичности генерации (0 – детерминированный вывод, 1 – максимально разнообразный вывод);
- **format** – требуемый формат ответа (например, "text" для обычного текста или "json" для структурированного вывода);
- **stream** – режим выдачи результата (при значении true ответ возвращается по мере генерации, при false – единым блоком) [15].

Правильное составление промпта имеет решающее значение для получения корректного ответа модели. Если запрос сформулирован нечетко или двусмысленно, даже самая мощная модель может выдать неверный или неуместный результат, что снизит качество работы всей системы [16]. В рамках прототипа особое внимание уделялось тому, чтобы промпт ясно описывал модельную задачу: например, содержал инструкции проанализировать конкретные данные и выдать ответ в требуемом формате (структурированном виде). Для удобства интеграции была использована официальная Python-библиотека Ollama [17], предоставляющий высокоуровневые функции для отправки запросов и получения ответов от локального сервера (см. рис. 1 и 2).

```
{
 "model": "llama3",
 "prompt": "Опишите роль нейросетей в современных производственных системах.",
 "temperature": 0.7,
 "format": "json",
 "stream": false
}
```

Рис. 1. Фрагмент запроса к Ollama

```python
import ollama
response = ollama.generate(
    model='llama3',
    prompt='Назовите ключевые принципы устойчивости нейронных сетей.',
    options={
        'temperature': 0.5,
        'format': 'json',
        'stream': False
    }
)
print(response['response'])
```

Рис. 2. Запрос к Ollama в Python

На рис. 3 представлен полученный результат, представляющий собой словарь, содержащий поля с метаданными, а также поле response, содержащее сгенерированный моделью текст.


```
{
 "model": "llama3",
 "created_at": "2025-03-24T12:34:56Z",
 "response": "Ключевыми принципами устойчивости нейронных
сетей являются способность к обобщению, толерантность к
шуму, адаптивность и интерпретируемость архитектуры.",
 "done": true
}
```


Рис. 3. Ответ модели

Кроме того, предусмотрена возможность включения дополнительных параметров, позволяющих более тонко настраивать поведение модели:

- top_p – параметр выборки по вероятностному порогу (nucleus sampling);
- num_ctx – максимальное количество токенов контекста;
- repeat_penalty – штраф за повторение одинаковых токенов;
- stop – список токенов-стопов, при достижении которых генерация прекращается [15].

4. ОБРАБОТКА МУЛЬТИМОДАЛЬНЫХ ДАННЫХ

Сервис построен по модульному принципу, где различные типы входных данных преобразуются в удобный для модели вид и объединяются в рамках единого запроса. Общая архитектура прототипа включает следующие компоненты:

- **видео:** периодически из видеопотока (IP-камеры наблюдения или видеозаписи) извлекаются кадры-изображения, которые затем могут быть поданы на вход модели;
- **сенсоры:** показания датчиков (например, температуры и влажности воздуха) агрегируются за небольшие интервалы времени и представляются в текстовом формате. Для эксперимента значения датчиков моделировались: были заданы нормальные и аномальные условия (повышение температуры, снижение влажности как индикатор возможного возгорания);
- **логи:** из внешних систем безопасности или контроля доступа берутся записи журнала событий за недавний промежуток времени. Эти текстовые записи включают отметки времени (в формате ISO 8601 [18]) и описание произошедших событий (например, срабатывание пожарной сигнализации, отключение датчика и т. д.). Для испытаний был подготовлен образец такого лога (например, фрагмент журнала системы контроля и управления доступом (СКУД)), пригодный для анализа моделью.

Все перечисленные выше данные формируются в единый промпт для модели. Таким образом, на вход модели поступает комплексная информация: одновременно и изображение с камеры, и соответствующие этому моменту показания сенсоров, и текстовые сообщения от других систем. Модель должна на основе всех вводных данных сформировать вывод о ситуации в наблюдаемой зоне. Благодаря использованию мультимодальных возможностей LLM (в частности, моделей, умеющих работать с визуальной информацией) вся аналитика выполняется единым интеллектуальным модулем – без необходимости раздельной обработки каждым источником и последующего объединения результатов. Это упрощает архитектуру и позволяет модели самой учитывать взаимосвязи между различными модальностями данных.

5. ЭКСПЕРИМЕНТАЛЬНАЯ МЕТОДИКА

Для проверки работоспособности прототипа и оценки эффективности разных моделей была разработана методика тестирования на основе нескольких сценариев. Общий процесс эксперимента состоит из трех этапов.

Этап 1. Подготовка тестовых данных. На первом этапе для каждой модальности были сформированы контрольные наборы данных, имитирующие ситуации в области наблюдения. В качестве видеоданных использовались изображения, сгенерированные нейросетью (модель OpenAI ChatGPT-4o-mini [19]), это позволило варьировать содержимое кадров (наличие людей, обстановка) и одновременно избежать использования реальных снимков. Для датчиков были заданы типичные ряды значений: в нормальных условиях – температура ~22 °C и влажность ~45%, в аномальном случае – резкое повышение температуры (до ~60 °C) и понижение влажности (< 20%) как признак возгорания. Кроме того, был подготовлен текстовый лог из системы безопасности: каждая запись содержала поле timestamp (время события) и поле event (описание самого события) (рис. 4). Такой лог имитировал внешние сигналы, дополняющие данные датчиков и видео.

Рис. 4. Сгенерированное фото с камеры видеонаблюдения. Человек упал. Зафиксирована аварийная ситуация.

Рис. 5. Сгенерированное фото с камеры видеонаблюдения. Пустой коридор в офисе. Система видеонаблюдения не обнаружила нарушений.

```
{
  "timestamp": "2025-05-15T12:00:00Z",
  "event": "Fire alarm triggered at Sector 7"
}
```

Рис. 6. Лог с системы безопасности. В данном примере поле timestamp означает время события в формате ISO 8601 [23], а event – описание самого события.

Этап 2. Выбор моделей и сценарии анализа. Для решения задачи были отобраны шесть мультимоделей, доступные в библиотеке Ollama: gemma3:12b [20], llava:13b [21], llama3.2-vision:11b [22], minicpm-v:8b [23], qwen2.5vl:7b [24], mistral-small3.2:24b [25]. Эти модели выбраны исходя из популярности и способности работать с изображениями наравне с текстом [14]. Каждая модель тестировалась на одном и том же наборе из четырех сценариев, заранее подготовленных на этапе 1. Каждый сценарий представлял собой комбинацию данных различных модальностей, соответствующих определенной ситуации.

Сценарий 1: «Человек упал». Видеокадр (условно рис. 4) содержит изображение человека, лежащего на полу без сознания; значения датчиков (рис. 6)

находятся в нормальном диапазоне (нет признаков возгорания или других аномалий). Ожидается, что модель, проанализировав картинку, распознает факт падения человека и сформирует вывод о критической ситуации (необходима помощь).

Сценарий 2: «Пожар с людьми». Камера зафиксировала в помещении присутствие людей (рис. 4, на изображении видны люди); датчики (рис. 7) показывают аномальные значения – высокая температура, низкая влажность; в логе присутствует запись о срабатывании пожарной сигнализации. Модель должна учесть все источники: по логам понять, что произошел пожар, по датчикам – подтверждение возгорания, по видео – наличие людей. Ожидаемый вывод: критическая ситуация, в помещении пожар и находятся люди, требуется немедленная реакция.

Сценарий 3: «Пожар без людей». На видеокадре (рис. 5) изображено пустое помещение или коридор; датчики (рис. 7) также сигнализируют о пожаре (высокая температура, сухость), но из логов нет сведений о присутствии людей. В этом случае модель должна сообщить о пожаре, подчеркнув отсутствие людей (тем не менее ситуация все равно критическая, требует вмешательства, например пожаротушения, но эвакуации людей не требуется).

Сценарий 4: «Штатный режим». Изображение камеры (рис. 5) – пустой коридор, все показатели датчиков в норме (используется тот же набор, что и в сценарии 1, рис. 6), внешних сигналов нет. Это контрольный сценарий благополучного состояния, на который модель не должна выдавать тревожную реакцию (ожидается, что система подтвердит отсутствие подозрительных событий).

Для автоматизации тестирования был написан скрипт на Python, который последовательно подставлял данные каждого сценария в промпт и опрашивал каждую из выбранных моделей через API Ollama. Скрипт измерял время выполнения запроса для каждой модели и сохранял ответы. Чтобы добиться воспроизводимых результатов, параметр temperature для моделей устанавливался равным 0 (детерминированная генерация), а формат ответа задавался как JSON с двумя полями: need_help (логический флаг, сигнализирует о необходимости реагирования) и message (текстовое описание ситуации от лица модели). Таким

образом, от каждой модели в каждом сценарии получался структурированный ответ (рис. 7).

```
{
  "need_help": true,
  "message": "Detected an unconscious person, emergency assistance required."
}
```

Рис. 7. Пример структурированного ответа от моделей.

Этап 3. Оценка результатов. На заключительном этапе проводилась оценка качества ответов моделей. Для каждой модели и каждого сценария заранее известен правильный ответ (требуется реакция или нет, корректное описание ситуации). Мы трактовали задачу как бинарную классификацию сценариев на требующие (критические) и нетребующие вмешательства (нормальные). На этой основе вычислялись стандартные метрики: точность (Precision), полнота (Recall) и сводная F1-мера для каждого набора ответов [26, 27]. Кроме того, для практической значимости сравнивалось среднее время отклика различных моделей. В табл. 1 приведены суммарные показатели качества классификации ситуаций для каждой модели, в табл. 2 – среднее время ответа модели на один сценарий.

5. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

По итогам экспериментов получены количественные оценки точности работы мультимодальных моделей в задаче мониторинга активности (выявления критических ситуаций). Табл. 1 демонстрирует сравнение методов с помощью метрик F1, Precision и Recall для шести моделей. Табл. 2 содержит данные о производительности – время, затрачиваемое моделями на обработку одного сценария в среднем (в секундах).

Табл. 1. Качество определения критической ситуации
(средние значения метрик по результатам 4-х сценариев)

Модель	Precision	Recall	F1-Score
gemma3:12b	1.00	0.67	0.80
llava:13b	0.00	0.00	0.00

minicpm-v:8b	1.00	0.50	0.67
qwen2.5vl:7b	0.00	0.00	0.00
mistral-small3.2:24b	1.00	0.67	0.80
llama3.2-vision:11b	1.00	0.50	0.67

Табл. 2. Время отклика моделей на один сценарий (в секундах)

Модель	Сценарий 1	Сценарий 2	Сценарий 3	Сценарий 4
gemma3:12b	17.06	6.42	5.89	6.11
llava:13b	21.74	12.20	10.61	9.32
minicpm-v:8b	5.87	1.84	3.61	1.49
qwen2.5vl:7b	20.41	18.20	17.97	18.37
mistral-small3.2:24b	43.54	40.89	37.45	37.91
llama3.2-vision:11b	31.92	29.33	28.70	29.91

Таким образом, из полученных результатов видно, лучшими оказались gemma3:12b и mistral-small3.2: они правильно отреагировали на три из четырех сценариев, что подтверждается наибольшим значением F1 = 0.8. Кроме того, они не допустили ни одного ложного срабатывания (Precision = 1.00), хотя и пропустили один из критических сценариев (Recall = 0.67). Модели minicpm-v:8b и llama3.2-vision:11b также продемонстрировали вполне приемлемую точность (F1 = 0.67), без ложных тревог, но с пропуском части инцидентов. Наихудший результат оказался у моделей llava:13b и qwen2.5vl:7b – они не смогли корректно идентифицировать ни одного сценария (все выходные ответы были ошибочными), о чем говорят нулевые значения метрик. Вероятно, модели оказались наименее подходящими для подобных комплексных запросов, возможно, из-за ограниченной специализации или недостаточной обучения для интеграции различных типов данных.

Что касается скорости работы, здесь лидирует облегченная модель minicpm-v:8b – ее среднее время отклика в простых сценариях составляло порядка 1.7 с (сценарии 2 и 4), и даже в более сложных ситуациях (сценарии 1 и 3) она укладывалась в 6 с. Модель gemma3:12b показывала стабильное время ответа около 10–20 с на сценарий, что быстрее тяжелых mistral-small3.2:24b и llama3.2-vision (на отдельных задачах время доходило до 45 с). Модели

llava:13b и qwen2.5vl:7b в целом работали сравнительно быстро (до 21 с), однако их низкие точности делают скорость несущественным фактором. Следует отметить, что все модели запускались на одной локальной машине, поэтому абсолютизировать приведенные цифры не стоит – при развертывании на производительном оборудовании время реакции может быть значительно снижено.

В целом эксперимент подтвердил возможность применения больших мультимодальных моделей для мониторинга: по крайней мере две из проверенных моделей (gemma3:12b и minicpm-v:8b) сумели обнаружить большинство заданных событий, правильно интерпретировав и совместив информацию из разных источников. Это весьма обнадеживающий результат, учитывая, что модели не проходили специального обучения под наши сценарии, а использовались «как есть». Таким образом, нейросети, предобученные на больших данных, в сочетании с грамотной инженерией промптов могут успешно решать задачи интеллектуального анализа ситуации.

Однако эксперимент выявил и ряд ограничений текущего прототипа. Во-первых, качество вывода значительно варьируется от модели к модели: выбор подходящей архитектуры критически влияет на точность. Модели, лучше настроенные на визуально-текстовый ввод (например, gemma3:12b), показали высокую результативность, тогда как другие оказались неприменимы в данном виде. Во-вторых, производительность системы пока оставляет желать лучшего – время отклика в десятки секунд неприемлемо для ряда практических сценариев (например, для систем реального времени, где счет может идти на секунды). Это частично связано с использованием больших моделей (12–13 млрд параметров) на CPU; ускорение возможно при переходе на GPU-версии или при оптимизации моделей (квантование, аппаратное ускорение). В-третьих, прототип был протестирован на ограниченном наборе синтетических данных. Отметим, что выбранные сценарии были приближены к реальным условиям, на практике могут возникать более сложные обстановки, шумы и непредусмотренные комбинации событий, где поведение модели потребует дополнительной проверки.

Тем не менее применимость в реальных условиях представляется вполне вероятной после доработки системы. Одним из преимуществ предложенного подхода является его гибкость: путем замены или обновления модели в Ollama

можно улучшить показатели без кардинальной переработки всей системы. Кроме того, локальное исполнение гарантирует, что чувствительные видеоданные и логи не покидают пределов устройства/локальной сети – это важно для организаций, предъявляющих строгие требования к безопасности данных (например, на производствах с режимом секретности или в медицинских учреждениях). Автономность решения означает и независимость от сетевой инфраструктуры: мониторинг не прервется даже при остановится доступа к Интернету или облаку.

ЗАКЛЮЧЕНИЕ

Разработанный прототип интеллектуального сервиса мультимодального мониторинга продемонстрировал перспективность применения больших предобученных нейросетевых моделей в системах отслеживания активности. В ходе экспериментов показано, что современные модели способны интегрированно анализировать данные разных типов (изображения, числовые показатели, текстовые события) и успешно выявлять сложные ситуации, ранее обнаруживаемые лишь человеком или узкоспециализированными алгоритмами. Использование локального фреймворка Ollama позволило запускать LLM-модели непосредственно на месте сбора данных, обеспечивая автономную работу системы и защиту информации. Полученные результаты подтверждают работоспособность подхода: наиболее точная модель (gemma3:12b) правильно распознала 75% сценариев, а более быстрая minicpm-v:8b при незначительном снижении полноты также может считаться успешной.

Вместе с тем проведенное исследование выявило и направления для дальнейшей работы. Одной из первоочередных задач является **повышение быстродействия**: планируется оптимизировать модели (например, за счет квантования до меньшей разрядности, использования версий «LoRA» или дистилляции знаний) и протестировать их на аппаратном ускорителе, чтобы добиться сокращения времени реакции до приемлемых величин. Еще одно перспективное направление – это обогащение интеллектуального анализа с помощью **онтологической поддержки сценариев**. Введение семантической модели предметной области (онтологии событий и объектов) могло бы помочь интерпретировать ответы модели и уменьшить вероятность ошибок, особенно в нетипичных случаях.

Кроме того, интеграция дополнительных модальностей (например, аудио, как обсуждалось выше) расширит возможности мониторинга: звук и речь могут предоставить ценные сведения о происходящем (крики, шумы аварий и пр.).

Настоящая работа выполнялась в рамках инициативного исследования, без прямого привлечения сторонних организаций. Первичное внедрение прототипа планируется осуществить на учебно-исследовательских стендах Казанского федерального университета, что позволит собрать дополнительную обратную связь и улучшить систему. В перспективе доработанное решение может быть опробовано в условиях, приближенных к промышленным, например в лабораториях или в рамках пилотного проекта на предприятии, заинтересованном в интеллектуальных системах безопасности. Таким образом, разработанный сервис представляет собой шаг вперед к созданию универсальных мультимодальных средств мониторинга, объединяющих достижения в области больших моделей с практическими требованиями автономности и безопасности.

СПИСОК ЛИТЕРАТУРЫ

1. *Onsu M.A., Lohan P., Kantarci B., Syed A., Andrews M., Kennedy S.* Leveraging Multimodal-LLMs Assisted by Instance Segmentation for Intelligent Traffic Monitoring [Электронный ресурс] // arXiv, 2025.
URL: https://arxiv.org/abs/2502.11304 (дата обращения: 15.05.2025).

2. *Ferrara E.* Large Language Models for Wearable Sensor-Based Human Activity Recognition, Health Monitoring, and Behavioral Modeling // Sensors. 2024. Vol. 24, No. 15. Article 5045.

3. *Suh S., Rey V.F., Lukowicz P.* Tasked: Transformer-based adversarial learning for human activity recognition using wearable sensors // Knowledge-Based Systems. 2023. Vol. 260. Article 110143.

4. *Миннеахметов Р. Р.* Интеллектуальный сервис мультимодального мониторинга области наблюдения // Научный сервис в сети Интернет: труды XXVI Всероссийской научной конференции (22–25 сентября 2025 г., онлайн). М.: ИПМ им. М.В. Келдыша, 2025.

5. *Nath N.D., Behzadan A.H., Paal S.G.* Deep learning for site safety: Real-time detection of personal protective equipment // Automation in Construction. 2020. Vol. 112. Article 103085.

6. *Gupta S.* Deep learning-based human activity recognition using wearable sensor data // International Journal of Information Management Data Insights. 2021. Vol. 1. Article 100046.

7. *Uçar A., Karakoşe M., Kırımça N.* Artificial Intelligence for Predictive Maintenance Applications: Key Components, Trustworthiness, and Future Trends // Applied Sciences. 2024. Vol. 14, No. 2. Article 898.

8. Bouchabou D., Nguyen S. M., Lohr C., LeDuc B., Kanellos I. A Survey of Human Activity Recognition in Smart Homes Based on IoT Sensors Algorithms: Taxonomies, Challenges, and Opportunities with Deep Learning // Sensors. 2021. Vol. 21. No. 18. Art. 6037. DOI: 10.3390/s21186037.

9. *Han S., Yuan S., Trabelsi M.* LogGPT: Log Anomaly Detection via GPT [Электронный ресурс] // arXiv. 2023. URL: https://arxiv.org/pdf/2309.14482 (дата обращения: 15.05.2025).

10. *Duong H.-T., Le V.-T., Hoang V. T.* Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey // Sensors. 2023. Vol. 23, No. 11. Art. 5024. DOI: 10.3390/s23115024.

11. *Özüağ S., Ertuğrul Ö.* Enhanced Occupational Safety in Agricultural Machinery Factories: Artificial Intelligence-Driven Helmet Detection Using Transfer Learning and Majority Voting // Applied Sciences. 2024. Vol. 14. Article 11278. https://doi.org/10.3390/app142311278

12. Radhi A., Altamimi Z., Dihin H., Husien W., Abbas O. A., Ahmed M. R. A., Hamad M., Al-Shimary A., Saleem H. Real-Time Human Activity Recognition in Smart Homes Using IoT Sensors and Deep Learning Models // Proceedings of the International Conference (ICBATS). 2025. P. 1–6. DOI: 10.1109/ICBATS66542.2025.11258537.

13. *Patel A. N., Murugan R., Maddikunta P. K. R., Yenduri G., Jhaveri R. H., Zhu Y., Gadekallu T. R.* AI-powered trustable and explainable fall detection system using transfer learning // Image and Vision Computing. 2024. Vol. 149. Art. 105164. DOI: 10.1016/j.imavis.2024.105164.

14. Ollama: [Электронный ресурс].

URL: https://ollama.com/ (дата обращения: 30.03.2025).

15. Ollama API Documentation: [Электронный ресурс]. URL: https://github.com/ollama/ollama/blob/main/docs/api.md (дата обращения: 30.03.2025).

16. *Sahoo P., Singh A.K., Saha S., Jain V., Mondal S., Chadha A.* A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications [Электронный ресурс] // arXiv. 2024. URL: https://arxiv.org/pdf/2402.07927 (дата обращения: 15.05.2025).

17. Ollama Python Library: [Электронный ресурс]. URL: https://github.com/ollama/ollama-python (дата обращения: 30.03.2025).

18. ISO 8601-1:2019 Standard: [Электронный ресурс]. URL: https://www.iso.org/obp/ui/#iso:std:iso:8601:-1:ed-1:v1:en (дата обращения: 30.03.2025).

19. OpenAI ChatGPT-4o-mini: [Электронный ресурс]. URL: https://chatgpt.com/ (дата обращения: 30.03.2025).

20. Ollama gemma3:12b Model: [Электронный ресурс]. URL: https://ollama.com/library/gemma3:12b (дата обращения: 30.03.2025).

21. Ollama llava:13b Model: [Электронный ресурс]. URL: https://ollama.com/library/llava:13b (дата обращения: 30.03.2025).

22. Ollama llama3.2-vision:11b Model: [Электронный ресурс]. URL: https://ollama.com/library/llama3.2-vision (дата обращения: 30.03.2025).

23. Ollama minicpm-v:8b Model: [Электронный ресурс]. URL: https://ollama.com/library/minicpm-v (дата обращения: 30.03.2025).

24. Ollama qwen2.5vl:7b Model: [Электронный ресурс]. URL: https://ollama.com/library/qwen2.5vl (дата обращения: 16.01.2026).

25. Ollama mistral-small3.2 Model: [Электронный ресурс]. URL: https://ollama.com/library/mistral-small3.2 (дата обращения: 16.01.2026).

26. *Hand D.J., Christen P.* F*: an interpretable transformation of the F-measure // Journal of Classification. 2021. Vol. 38, No. 1. P. 3–17.

27. Scikit Learn F1-Score: [Электронный ресурс]. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html (дата обращения: 30.03.2025).

INTELLIGENT MULTIMODAL NEURAL NETWORK MONITORING SERVICE FOR THE SURVEILLANCE AREA

R. R. Minneakmetov[0009-0007-8551-1393]

Kazan (Volga Region) Federal University, Kazan, Russia

razil0071999@gmail.com

Abstract

The article presents an approach to the development of an intelligent multimodal monitoring service for the surveillance area using large neural network models. The proposed solution is capable of analyzing heterogeneous data – video streams, environmental sensor signals (temperature, humidity, etc.), and event logs – to obtain a complete picture of what is happening. The main tools used are large language and visual models (for example, LLaMA, MiniCPM-V, etc.) deployed locally using the Ollama platform, which provides autonomous and secure information processing without the need to transfer data to the cloud. A prototype system has been developed that works offline and is capable of detecting critical situations, abnormal deviations from the norm and contextually significant events in the observed area. The method of forming test scenarios and conducting a qualitative assessment of the model's performance using the metrics F1-measure, Precision, Recall on a set of various situations is described. The experimental results confirm the applicability of multimodal models for monitoring tasks: the prototype successfully recognizes complex patterns of behavior and demonstrates the potential of large models in building adaptive and scalable surveillance systems.

Keywords*: intelligent service, multimodal monitoring, Ollama, Large Language Models, activity tracking, video analytics, artificial intelligence.*

REFERENCES

1. *Onsu M.A., Lohan P., Kantarci B., Syed A., Andrews M., Kennedy S.* Leveraging Multimodal Large Language Models Assisted by Instance Segmentation for

Intelligent Traffic Monitoring [Electronic resource] // arXiv. 2025. Available at: https://arxiv.org/abs/2502.11304 (accessed: 15.05.2025).

2. *Ferrara E.* Large Language Models for Wearable Sensor-Based Human Activity Recognition, Health Monitoring, and Behavioral Modeling // Sensors. 2024. Vol. 24, No. 15. Article 5045.

3. *Suh S., Rey V.F., Lukowicz P.* Tasked: Transformer-Based Adversarial Learning for Human Activity Recognition Using Wearable Sensors // Knowledge-Based Systems. 2023. Vol. 260. Article 110143.

4. Minneakhmetov R. Intelligent multimodal monitoring service for the surveillance area // All-Russian Conference "Scientific Services & Internet" (September 22–25, 2025, online). Moscow. Keldysh Institute of Applied Mathematics. 2025.

5. *Nath N.D., Behzadan A.H., Paal S.G.* Deep Learning for Site Safety: Real-Time Detection of Personal Protective Equipment // Automation in Construction. 2020. Vol. 112. Article 103085.

6. *Gupta S.* Deep Learning-Based Human Activity Recognition Using Wearable Sensor Data // International Journal of Information Management Data Insights. 2021. Vol. 1. Article 100046.

7. *Uçar A., Karakoşe M., Kırımça N.* Artificial Intelligence for Predictive Maintenance Applications: Key Components, Trustworthiness, and Future Trends // Applied Sciences. 2024. Vol. 14, No. 2. Article 898.

8. Bouchabou D., Nguyen S. M., Lohr C., LeDuc B., Kanellos I. A Survey of Human Activity Recognition in Smart Homes Based on IoT Sensors Algorithms: Taxonomies, Challenges, and Opportunities with Deep Learning // Sensors. 2021. Vol. 21. No. 18. Art. 6037. DOI: 10.3390/s21186037.

9. *Han S., Yuan S., Trabelsi M.* LogGPT: Log Anomaly Detection via GPT [Electronic resource] // arXiv. 2023. Available at: https://arxiv.org/pdf/2309.14482 (accessed: 15.05.2025).

10. *Duong H.-T., Le V.-T., Hoang V. T.* Deep Learning-Based Anomaly Detection in Video Surveillance: A Survey // Sensors. 2023. Vol. 23, No. 11. Art. 5024. DOI: 10.3390/s23115024.

11. *Özüağ S., Ertuğrul Ö.* Enhanced Occupational Safety in Agricultural Machinery Factories: Artificial Intelligence-Driven Helmet Detection Using Transfer Learning and Majority Voting // Applied Sciences. 2024. Vol. 14. Article 11278. https://doi.org/10.3390/app142311278.

12. *Radhi A., Altamimi Z., Dihin H., Husien W., Abbas O. A., Ahmed M. R. A., Hamad M., Al-Shimary A., Saleem H.* Real-Time Human Activity Recognition in Smart Homes Using IoT Sensors and Deep Learning Models // Proceedings of the International Conference (ICBATS). 2025. P. 1-6. DOI: 10.1109/ICBATS66542.2025.11258537.

13. *Patel A. N., Murugan R., Maddikunta P. K. R., Yenduri G., Jhaveri R. H., Zhu Y., Gadekallu T. R.* AI-powered trustable and explainable fall detection system using transfer learning // Image and Vision Computing. 2024. Vol. 149. Art. 105164. DOI: 10.1016/j.imavis.2024.105164.

14. Ollama [Electronic resource]. Available at: https://ollama.com/ (accessed: 30.03.2025).

15. Ollama API Documentation [Electronic resource]. Available at: https://github.com/ollama/ollama/blob/main/docs/api.md (accessed: 30.03.2025).

16. *Sahoo P., Singh A.K., Saha S., Jain V., Mondal S., Chadha A.* A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications [Electronic resource] // arXiv. 2024. Available at: https://arxiv.org/pdf/2402.07927 (accessed: 15.05.2025).

17. Ollama Python Library [Electronic resource]. Available at: https://github.com/ollama/ollama-python (accessed: 30.03.2025).

18. ISO 8601-1:2019 Standard [Electronic resource]. Available at: https://www.iso.org/obp/ui/#iso:std:iso:8601:-1:ed-1:v1:en (accessed: 30.03.2025).

19. OpenAI ChatGPT-4o-mini [Electronic resource]. Available at: https://chatgpt.com/ (accessed: 30.03.2025).

20. Ollama Gemma3:12B Model [Electronic resource]. Available at: https://ollama.com/library/gemma3:12b (accessed: 30.03.2025).

21. Ollama LLaVA:13B Model [Electronic resource]. Available at: https://ollama.com/library/llava:13b (accessed: 30.03.2025).

22. Ollama Llama3.2-Vision:11B Model [Electronic resource]. Available at: https://ollama.com/library/llama3.2-vision (accessed: 30.03.2025).

23. Ollama MiniCPM-V:8B Model [Electronic resource]. Available at: https://ollama.com/library/minicpm-v (accessed: 30.03.2025).

24. Ollama Qwen2.5-VL:7B Model [Electronic resource]. Available at: https://ollama.com/library/qwen2.5vl (accessed: 16.01.2026).

25. Ollama Mistral-Small-3.2 Model [Electronic resource]. Available at: https://ollama.com/library/mistral-small3.2 (accessed: 16.01.2026).

26. *Hand D.J., Christen P.* F*: An Interpretable Transformation of the Measure // Journal of Classification. 2021. Vol. 38, No. 1. P. 3–17.

27. Scikit-learn F1-Score [Electronic resource]. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html (accessed: 30.03.2025).

СВЕДЕНИЯ ОБ АВТОРЕ

МИННЕАХМЕТОВ Разиль Рустемович – магистр программной инженерии, аспирант Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета.

Razil MINNEAKHMETOV – Master of Software Engineering, PhD student at the Institute of Information Technology and Intelligent Systems, Kazan (Volga Region) Federal University. Current scientific interests: artificial intelligence, large neural models, recommender systems, cloud computing, internet of things.

email: razil0071999@gmail.com

ORCID: 0009-0007-8551-1393

Материал поступил в редакцию 16 января 2026 года