Упрощенный метод контролируемого обнаружения видеоаномалий с извлечением на уровне кадров на датасете UCF-Crime

К. М. Мостафа[1], М. А. Мовафи[2], С. А. Аббас[3], К. С. Аттиа[4]

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

[1]kareem50329@gmail.com, [2]mowafy7@yandex.ru
[3]saddamabbas077@gmail.com, [4]attiakarim2001@gmail.com

Аннотация. **Обнаружение аномальных событий в видеопотоках с камер наблюдения является важной задачей для автоматизации систем мониторинга безопасности. Тем не менее, данная задача остаётся сложной из-за недостатка покадровой разметки и высоких вычислительных затрат, необходимых для обработки длительных видеопоследовательностей. В настоящей статье предлагается вычислительно эффективная платформа для слабо контролируемого обнаружения видеоаномалий, которая позволяет определять временную локализацию и выделять аномальные события на уровне кадров, используя только метки на уровне видео. Предложенная система оценивается на наборе данных UCF-Crime и рассчитана на работу на оборудовании потребительского класса без привлечения крупномасштабных вычислительных ресурсов. В статье представлен двухэтапный конвейер. На первом этапе автоэнкодер последовательностей на основе BiLSTM, обученный исключительно на нормальных (неаномальных) видео, вычисляет ошибку восстановления видеосегментов и выступает в роли детектора аномалий. На втором этапе сеть SlowFast-R50 выполняет классификацию типов преступлений, используя как медленный, так и быстрый временные потоки. Кроме того, предлагается метод выбора видеосегментов с учётом движения, основанный на вычислении оптического потока для всех кадров, что гарантирует, что каждый клип, подаваемый на вход конвейера, действительно содержит событие, а не относится к интервалу до или после его совершения. Выполнено сравнение двух вариантов конвейера. В первом варианте в медленный поток направляются все сегменты криминального видео независимо от выходного сигнала вентильного фильтра (gate). Во втором варианте в медленный поток передаются только те клипы, которые прошли через вентильный фильтр.**

Ключевые слова: обнаружение видео-аномалий; обучение со слабым контролем; обучение с несколькими экземплярами (MIL); автоэнкодер LSTM; медленная сеть; оптический поток; анализ видео наблюдения; глубокое извлечение признаков; сверточные нейронные сети; классификация на уровне сегмента.

I. Введение

Анализ видеозаписей с камер наблюдения представляет собой популярное направление исследований, имеющее практическое значение в таких областях, как общественная безопасность и раскрытие преступлений, и может быть использован для создания инфраструктуры «умного города». Несмотря на многочисленные достижения в применении глубокого обучения к видеоданным, реальное внедрение этих методов в практику раскрытия преступлений остаётся сложной задачей в силу двух основных препятствий: редкого присутствия противоправных действий по сравнению с фоновыми кадрами и высокого сходства различных типов преступлений, обусловленного их пространственным проявлением.

В традиционных методах обнаружения аномалий в видеоклипах использовались ручные подходы к извлечению признаков, такие как гистограмма ориентированных градиентов (HOG) и гистограммы оптического потока, для применения в рамках классификации одним классом и методов разреженного кодирования. Хотя эти подходы заложили основы для данной области, они показывают низкую эффективность при изменяющемся освещении и движении камеры, особенно при работе с реалистичными видеороликами. Современные решения данной проблемы включают использование свёрточных нейронных сетей (CNN) и рекуррентных нейронных сетей для моделирования внешнего вида и временных изменений в видео.

Указанные ограничения преодолеваются с помощью единого подхода к проектированию, который использует:

1) обрезку видео с учётом движения, управляемую оптическим потоком, для выделения окон с максимальной активностью;

2) автоэнкодер последовательностей на основе BiLSTM, обученный исключительно на нормальных клипах и выступающий в роли детектора аномалий, учитывающего временную динамику;

3) обучение классификатора преступлений SlowFast R50 на клипах, проходящих через вентильный фильтр аномалий.

Сравниваются два подхода. В первом подходе все криминальные клипы участвуют в обучении классификатора преступлений. Во втором подходе классификатор обучается только на тех клипах, которые прошли контроль аномалий.

II. Обзор сопутствующих работ

Задача обнаружения видеоаномалий широко исследовалась в рамках одноклассовой постановки (one-class classification), при которой для обучения предоставляются только примеры, принадлежащие к рассматриваемому классу. В работе Hasan и другие. [1] был разработан свёрточный автоэнкодер для

обнаружения аномалий путём обучения нормальному восстановлению кадров; при этом большие ошибки реконструкции служат индикаторами аномалий. Лю и другие. [2] расширили данный подход, предложив предсказание будущих кадров с помощью архитектур на основе U-Net. Авторы показали, что ошибки прогнозирования во временных кадрах являются более надёжными индикаторами аномалий по сравнению с ошибками реконструкции в статических кадрах.

Широкое распространение также получили графовые методы и подходы, основанные на механизме внимания. Чжун и другие. [3] сформулировали задачу обнаружения аномалий как слабо контролируемую обучающую задачу, в которой используются метки на уровне видео, и предложили метод временного ранжирования на наборе данных UCF-Crime. В своём подходе к обнаружению аномалий Ли и другие. [4] применили трансформеры для встраивания механизмов самовнимания поверх фрагментов кадров, что позволило фиксировать структуры зависимостей, значимые для оценки аномалий. В настоящем же исследовании мы использовали архитектуру BiLSTM для моделирования последовательности признаков кадров, зависящей от времени, получаемых из свёрточной нейронной сети (CNN).

Помимо обнаружения аномалий, классификация типов преступлений также привлекла значительное внимание после публикации набора данных UCF-Crime [5], предложенного Султани и другие, который содержит 1900 видеозаписей, отнесённых к 13 классам преступлений. Тянь и другие. [6] предложили использовать слабо контролируемое обучение с несколькими экземплярами (MIL) для распознавания типов преступлений без необходимости разметки каждого кадра, что позволило достичь высоких базовых показателей на наборе данных UCF-Crime. Позднее Ву и другие. [7] применили графовые свёрточные сети для изучения взаимосвязей между обнаруженными объектами в видеокадрах, повышая точность распознавания подтипов преступлений.

В последнее время модели видеотрансформеров, такие как TimeSformer [8] и Video Swin Transformer [9], достигли высокой производительности в задачах распознавания действий на наборах данных, включая Kinetics-400 и Something-Something. Тем не менее, такие модели требуют значительных вычислительных ресурсов, что ограничивает их применение в реальных условиях. Сеть SlowFast [10], использующая два потока на основе остаточных соединений с различной временной дискретизацией для извлечения семантической информации о внешнем виде и движении, представляет собой практичную альтернативу для решения указанной проблемы. Поэтому в нашей работе используется сеть SlowFast-R50 в качестве классификатора преступлений ввиду её эффективности при распознавании действий, зависящих от движения.

Мостафа и др. [11] представляют модель, схожую с этим исследованием, предлагая двухэтапную структуру для извлечения ключевых кадров из видео со слабым наблюдением и детальной категоризацией преступлений, используя тот же набор данных наблюдения 13-го класса, полученный из набора данных UCF-Crime dataset, который использовался в нашем эксперименте.

III. Предлагаемая методология

Конвейер включает три модуля: модуль извлечения клипов с учётом движения, вентильный детектор аномалий на основе BiLSTM и классификатор преступлений SlowFast. Общая архитектура модели представлена на схеме ниже.

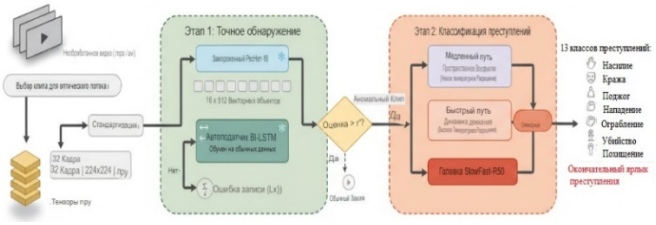

Рис. 1. Рабочий процесс предлагаемого способа

A. *Модуль извлечения клипов*

Основным недостатком предшествующих исследований являлся способ временной дискретизации, используемый для формирования обучающих клипов из очень длинных видеозаписей наблюдения. Проблема заключалась в том, что криминальные события происходили редко в рамках столь протяжённых видеороликов, и поэтому клипы, отобранные случайным образом, с высокой вероятностью содержали кадры, которые напоминали легитимные (обычные) кадры либо до, либо после совершения преступления. Это приводило к появлению шума в данных, несмотря на корректную маркировку на уровне самих видеороликов.

Для решения данной проблемы была выполнена оценка каждого возможного окна длиной в T кадров внутри видео на основе величины плотного оптического потока, используемой в качестве показателя активности. Процесс вычисления оценок заключается в расчёте средней величины оптического потока (метод Farneback) между каждой парой последовательных кадров в окне. Каждая пара кадров для повышения эффективности дискретизируется до разрешения 112×112 пикселей. Далее окна упорядочиваются по полученным оценкам, после чего отбираются верхние K непересекающихся окон с установлением минимальной временной разницы между любыми двумя кадрами. Затем применяется пороговое значение движения для исключения статичных окон. Для видеороликов обычного (неаномального) класса необходима единообразная стратегия выборки с целью предотвращения разнообразия сцен, обусловленного движением.

B. *Шлюз аномалий BiLSTM*

Вентильный детектор аномалий (anomaly gate) представляет собой одноклассовый классификатор, обученный исключительно на нормальных (неаномальных) клипах. Он предназначен для определения того, является ли видеоклип аномальным, при этом использует менее сложную модель по сравнению с сетью SlowFast.

1) Извлечение объектов: Модель ResNet-18, предварительно обученная на наборе данных ImageNet, применяется для извлечения признаков каждого кадра в виде 512-мерных векторов. Из видеоклипа, содержащего 32 кадра, равномерно выбираются 16 кадров, для которых извлекаются векторы признаков.

2) Архитектура автоэнкодера: Модель вентильного детектора использует автоэнкодер последовательностей на основе BiLSTM. Кодировщик представляет собой двунаправленную сеть BiLSTM с двумя уровнями и размером скрытого состояния 256 в каждом направлении. Объединение прямого и обратного скрытых состояний последнего уровня кодировщика проецируется через линейное узкое место (бутылочное горлышко) в скрытое пространство размерности 256. Декодер представляет собой однослойный однонаправленный LSTM, который принимает на входе вектор скрытого узкого места, повторяющийся T раз, и пытается восстановить исходную последовательность входных признаков. В качестве функции потерь при обучении используется среднеквадратичная ошибка (MSE).

3) Оценка аномалий: На этапе вывода ошибка восстановления для отдельных выборок, вычисляемая как MSE между входной последовательностью и её реконструкцией, нормируется к диапазону [0, 1] с использованием калибровочного коэффициента. Этот коэффициент основан на 95-м процентиле ошибок восстановления, наблюдаемых на нормальных выборках, не участвовавших в обучении. Пороговое значение для вентильного детектора выбирается путём максимизации индекса Юдена (чувствительность + специфичность − 1, или истинно положительная частота минус ложно положительная частота) на ROC-кривой валидационной выборки.

C. Медленный классификатор

Если входной видеоклип был идентифицирован как аномальный, модель SlowFast-R50 обеспечивает точную классификацию с высокой детализацией, относя его к одному из 13 различных типов преступлений. Архитектура SlowFast использует два боковых пути (два потока), каждый из которых содержит сетевую архитектуру ResNet-50. Медленный путь (slow path) осуществляет дискретизацию последовательности с низкой временно́й частотой, выбирая один кадр из четырёх в последовательности из 32 кадров, что позволяет захватывать контекстную информацию о сцене. Быстрый путь (fast path) использует все 32 кадра, но с уменьшенным в восемь раз количеством каналов по сравнению с медленным путём, что предназначено для захвата динамики движений. Исходный классификационный слой, предназначенный для набора данных Kinetics-400, заменён на слой, специализированный для распознавания типов преступлений.

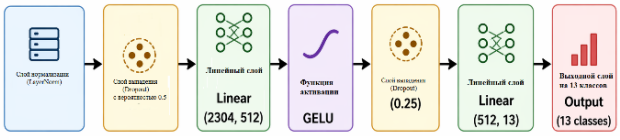

Рис. 2. Классификационная головка

В работе используется оптимизатор AdamW с двумя группами параметров. Для магистральной сети (backbone) SlowFast скорость обучения установлена в 10 раз меньше базовой и составляет lr = 1×10^{-5}, тогда как для новой классификационной головки используется базовая скорость обучения lr = 1×10^{-4}. Применяется планировщик скорости обучения с одноцикловым режимом (one-cycle), включающий 10-процентный этап прогрева и график косинусного отжига (cosine annealing). Для устранения дисбаланса классов используются взвешенная случайная выборка (weighted random sampling) и обратное частотное взвешивание функции потерь кросс-энтропии; параметр сглаживания меток (label smoothing) установлен равным ε = 0,1. Методы аугментации данных включают случайные горизонтальные отражения (с вероятностью 0,5) и цветовое дрожание (colour jitter) с вероятностью 0,3.

D. Полный конвейер

Наиболее существенное архитектурное различие между двумя рассматриваемыми системами заключается в том, какие именно видеозаписи криминального характера подаются на вход сети SlowFast в процессе обучения. В первой модели любые видеоролики, содержащие преступные деяния, могут быть переданы в сеть SlowFast. Во второй модели, напротив, в сеть SlowFast направляются только те видеоролики, для которых значение аномалии превысило установленный порог вентильного детектора (gate). Во избежание недостатка данных в каждой категории предусмотрено наличие не менее 20 видеороликов на класс. Сравнение архитектурных различий представлено в табл. 1.

ТАБЛИЦА I. СРАВНЕНИЕ АРХИТЕКТУРНЫХ РАЗЛИЧИЙ

Аспект	Первая модель (без фильтра)	Вторая модель (фильтр gate)
Обучающие данные	Только для всех клипов	криминальные клипы
Охват клипов	100% криминальных клипов	Зависит от gate (~ 70%)
Нехватка данных	Неприменимо	защита пола (20)
Прочность	Максимальный объем данных	Более чистое распределение
Ожидаемый риск	Клипы вне зоны доступа	Меньшее количество выборок (ограничения доступа)

IV. ЭКСПЕРИМЕНТЫ И РЕЗУЛЬТАТЫ

Эксперименты проводились с использованием набора данных, содержащего 14 классов, из которых 13 классов относятся к категории преступлений, а один класс является нормальным (обычным). Данный набор данных включает примерно 1000 длинных видеороликов. Применение метода извлечения клипов с учётом движения с параметром K = 6 клипов на видео позволило извлечь из всего набора данных около 5557 клипов, относящихся к преступлениям, и 300 нормальных клипов.

Эксперименты выполнялись с использованием одного графического процессора NVIDIA, оснащённого 16 ГБ видеопамяти. Обучение вентильного детектора (gate) проводилось в течение 40 эпох с размером пакета (batch size), равным 16. При этом использовался планировщик скорости обучения ReduceLROnPlateau. Обучение модели SlowFast осуществлялось в течение 20 эпох с размером пакета, равным 4.

A. Ворота

Автоэнкодер на основе BiLSTM, обученный в рамках обоих конвейеров, достигает значения ROC-AUC, равного 0,74, что является высоким показателем для вычислительно облегчённой модели. Оптимизированный порог, определённый по индексу

Юдена (Youden's J), успешно блокирует доступ 68,9% нормальных видеоклипов и пропускает 69,8% аномальных видеоклипов, содержащих сцены преступлений. Данный результат объясняется способностью автоэнкодера BiLSTM моделировать временную структуру последовательностей кадров при их восстановлении.

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}\big(\mathrm{FPR}^{-1}(u)\big)\, du \approx \sum_{k=1}^{K-1} \frac{(\mathrm{FPR}_{k+1} - \mathrm{FPR}_k)(\mathrm{TPR}_{k+1} + \mathrm{TPR}_k)}{2} \tag{1}$$

B. Медленный классификатор

Первый конвейер (в котором используются все криминальные видеоролики без фильтрации через вентильный детектор) достигает точности классификации преступлений 98,1% на валидационной выборке, отфильтрованной с помощью вентильного детектора, при значении функции потерь 0,82. Применение планировщика OneCycleLR и сглаживания меток (label smoothing) вносит существенный вклад в стабильность сходимости модели.

Второй конвейер (в котором обучающие видеоролики проходят фильтрацию через вентильный детектор) обеспечивает точность классификации преступлений 96,9% при значении функции потерь 1,01. Количество обучающих клипов составляет около 3800, что соответствует 68,3% от исходных 5557 клипов, поскольку остальные клипы были отфильтрованы. Уменьшение обучающей выборки само по себе не приводит к значимым изменениям, однако подтверждает предположение о том, что согласование распределения обучающих данных с распределением данных, используемых на этапе вывода, способно повысить точность классификатора.

$$\mathrm{Acc}_{\text{classifier}} = \frac{1}{N_{\text{val}}^{\text{crime}}} \sum_{i=1}^{N_{\text{val}}^{\text{crime}}} \mathbf{1}\left[\arg\max_k \; p_k^{(i)} = y_i\right] \tag{2}$$

C. Сравнительный анализ

Было проведено сравнение предложенного метода с современными аналогами. Разработанный подход демонстрирует улучшение в решении задачи классификации видеозаписей преступлений, которая являлась основной целью исследования. Нормальные видеоролики, используемые для распознавания событий, были направлены на поддержку процесса обучения и повышение качества классификации.

ТАБЛИЦА II.

Способ	***Достоверность***	***Потери***
конечный метод производство (все клипы)	96.9%	0.82
конечный способ трубопровода (фильтрованный клипы)	98.1%	1.01
предложенный метод	85.7 %	0

D. Ограничения и случаи отказа

Предложенная модель продемонстрировала высокую точность: все категории преступлений показывают многообещающие характеристики, что позволяет поддерживать интеграцию в реальном времени для приложений в области общественной безопасности. Тем не менее, вентильный детектор (gate), используемый для точного обнаружения аномальной активности, направляет любые ложные признаки преступной деятельности в классификатор. Это приводит к ошибочному отнесению некоторых нормальных событий к категориям преступлений.

V. Заключение

В данной работе представлены два основных вклада в разработку двухэтапной системы обнаружения и классификации преступлений на основе видеозаписей с камер наблюдения с учётом движения. Эти вклады включают: (1) плотное извлечение клипов на основе оптического потока для генерации истинных окон активности; (2) автоэнкодер последовательностей на основе BiLSTM для обнаружения аномалий путём использования ошибки временной реконструкции; (3) сравнительный анализ схем обучения классификатора — не зависящих от вентильного детектора и согласованных с ним.

Вентильный детектор на основе BiLSTM продемонстрировал успешную работу (ROC-AUC = 0,74) со следующими показателями: корректное блокирование нормальных событий — 68,9%, корректное обнаружение аномалий — 69,8%. При использовании двухканального классификатора SlowFast-R50, обученного в согласованном с вентильным детектором режиме, конвейер достиг точности 96,9% при значении функции потерь 1,01. В то же время конвейер без предварительной фильтрации обучающих видеороликов показал точность 98,1% и значение функции потерь 0,82. Данный результат подчёркивает важность выравнивания распределений данных на этапе логического вывода в каскадных системах анализа видео.

Список литературы

[1] М. Хасан, Дж. Чой, Дж. Нейман, А. К. Рой-Чоудхури и Л.С. Дэвис, "Изучение временной регулярности в видеопоследовательностях", в Proc. CVPR, 2016, с. 733-742.

[2] У. Лю, У. Ло, Д. Лиан и С. Гао, "Прогнозирование кадра будущего для обнаружения аномалий – новый базовый уровень", в Proc. CVPR, 2018.

[3] Дж. Чжун, Н. Ли, У. Конг, С. Лю, Т. Х. Ли и Г. Ли, "Очиститель шума сверточных меток графа: обучите классификатор действий plug-and-play для обнаружения аномалий", в Proc. CVPR, 2019.

[4] С. Ли, Ф. Феng, Л. Ван, А. Улла, А. Доерманн и М. Элосейни, "Обнаружение аномалий на основе трансформатора для видео наблюдения", arXiv: 2112.06830, 2021.

[5] У. Султани, К. Чен и М. Шах, "Обнаружение аномалий в реальном мире на видео наблюдения", в Proc. CVPR, 2018, с. 6479-6488.

[6] Ю. Тянь, Г. Панг, Ю. Чен, Р. Сингх, Дж. У. Верджанс и Г. Карнейро, "Обнаружение видео-аномалий со слабым контролем с надежным изучением величины временных признаков", в Proc. ICCV, 2021.

[7] П. Ву, Дж. Лю, Ю. Ши, Ю. Сун, Ф. Шао, З. Ву и З. Янг, "Не только смотрите, но и слушайте: учимся выявлять мультимодальное насилие под слабым наблюдением", в Proc. ECCV, 2020.

[8] Г. Бертасиус, Х. Ванг и Л. Торресани, "Является ли пространственно-временное внимание всем, что вам нужно для понимания видео?" в Proc. ICML, 2021.

[9] З. Лю, Дж. Нин, Ю. Цао, Ю. Вэй, З. Чжан, С. Линь и Х. Ху, "Видео-трансформер swin", в Proc. CVPR, 2022, с. 3202-3211.

[10] К. Фейхтенхофер, Х. Фан, Дж. Малик и К. Он, "Медленные сети для распознавания видео", в Proc. ICCV, 2019, с. 6202-6211.

[11] Мостафа К.М., Мохаммед А.С., Аббас С.А. и Аттия, К.С. (2025). Использование методов видеонаблюдения и глубокого обучения для обеспечения максимальной точности в приложениях для обеспечения общественной безопасности. 2025 VI Международная конференция по нейронным сетям и нейротехнологиям (NeuroNT), 83-85.