УДК 004.932:004.85

Н.А. Бессонов*
студент
*Санкт-Петербургский государственный университет аэрокосмического приборостроения

ПРИМЕНЕНИЕ АЛГОРИТМА SLOWFAST ДЛЯ АНАЛИЗА ВИДЕОДАННЫХ В ЗАДАЧАХ РАСПОЗНАВАНИЯ ДЕЙСТВИЙ ЧЕЛОВЕКА В РЕАЛЬНОМ ВРЕМЕНИ

Аннотация

Рассматривается задача распознавания действий человека в реальном времени на основе архитектуры SlowFast. Представлена математическая модель двухпутевой обработки видеопотока с разделением на медленный и быстрый каналы, описаны этапы работы алгоритма и особенности его применения для анализа высокоскоростных видеоданных. Приведены примеры практического использования метода в системах видеонаблюдения, интеллектуального управления и промышленной безопасности.

Ключевые слова: распознавание действий, SlowFast, видеоаналитика, реальное время, глубокое обучение, двухпоточные сети, пространственно-временное моделирование.

Abstract

The paper considers the application of the SlowFast neural network architecture for video data analysis in real-time human action recognition systems. The mathematical foundation of the two-pathway architecture is presented, including the separation of spatial and temporal processing. Advantages and limitations of the method when processing high-frame-rate video streams are analyzed. Practical examples of using SlowFast in video surveillance, human–computer interaction, and industrial safety systems are discussed.

Keywords: SlowFast, human action recognition, video analysis, real-time systems, deep learning, two-stream architecture, spatiotemporal modeling.

Введение

В условиях цифровой трансформации общества видеоданные становятся одним из наиболее информативных и быстрорастущих типов информационных ресурсов. Системы видеонаблюдения, камеры мобильных устройств, платформы потокового видео и промышленные сенсоры ежедневно генерируют петабайты видеопотоков. Эффективный анализ таких данных, особенно в реальном времени, требует применения интеллектуальных методов компьютерного зрения. Современные организации сталкиваются с необходимостью оперативного распознавания действий человека - от бытовых жестов до сложных производственных операций. В этом контексте системы анализа видеоданных, использующие глубокое обучение, становятся ключевым инструментом обеспечения безопасности, автоматизации процессов и создания адаптивных интерфейсов [1].

Одной из ключевых задач является распознавание действий человека в реальном времени - процесс классификации видеопоследовательностей с целью определения выполняемого действия (ходьба, падение, поднятие предмета, взаимодействие с оборудованием). Решение этой задачи позволяет повысить эффективность систем видеонаблюдения за счет автоматического обнаружения инцидентов, улучшить качество человеко-машинного взаимодействия, снизить риски на производстве и создать основу для интеллектуальных систем помощи операторам [2].

Для решения данной задачи широко применяются методы глубокого обучения, в частности архитектуры на основе сверточных нейронных сетей (CNN), адаптированные для

работы с видео. Одной из наиболее эффективных и получивших широкое признание архитектур является SlowFast, предложенная исследователями из Facebook AI Research (FAIR) в 2019 году [3]. Ее популярность обусловлена балансом между вычислительной эффективностью и высокой точностью распознавания. SlowFast имитирует работу зрительной коры млекопитающих, обрабатывающих движение на разных скоростях, что позволяет одновременно улавливать статический контекст (объекты, сцены) и быстрые движения (жесты, перемещения). Алгоритм успешно применяется как в исследовательских проектах (Kinetics, AVA), так и в промышленных системах реального времени.

Теоретические основы алгоритма SlowFast

Архитектура SlowFast относится к классу двухпутевых (two-pathway) моделей для анализа видео. Основная идея заключается в использовании двух параллельных ветвей: медленной (Slow pathway) и быстрой (Fast pathway). Медленная ветвь обрабатывает видео с низким темпом сэмплирования кадров (обычно 1-2 кадра в секунду) и захватывает семантическую информацию о сцене и объектах. Быстрая ветвь работает с высоким темпом сэмплирования (частота в 4-8 раз выше) и имеет уменьшенную пространственную размерность, что позволяет ей фокусироваться на временных изменениях и быстрых движениях.

Математически модель можно описать следующим образом. Пусть входной видеоклип представляет собой тензор $V \in R\hat{}\{T * H * W * 3\}$, где T - число кадров, H, W - высота и ширина, 3 - цветовые каналы. Медленная ветвь сэмплирует кадры с шагом τ (например, $\tau = 16$), быстрая - с шагом τ/α, где $\alpha > 1$ (обычно $\alpha = 8$). Пространственное разрешение быстрой ветви уменьшается в β раз (обычно $\beta = 1/8$ от разрешения медленной ветви). Функция потерь для обучения является комбинацией стандартной кросс-энтропии для классификации действий и дополнительных потерь, обеспечивающих согласование признаков между ветвями:

$L = L_{cls} + \lambda L_{lateral}$,

где L_{cls} - классификационная потеря, $L_{lateral}$ - потеря согласования через латеральные соединения (lateral connections), передающие информацию от быстрой ветви к медленной. Латеральные соединения реализуются через операции свертки и объединения, выравнивающие размерности признаков [3].

Алгоритм начинает работу с инициализации обеих ветвей предобученными на изображениях сверточными сетями (например, ResNet или ResNeXt). Затем выполняется совместное сквозное обучение на видеоданных с размеченными действиями. После обучения для нового видеоклипа вычисляются признаки обеих ветвей, объединяются (обычно конкатенацией) и подаются в классификатор, который выдает вероятности классов действий. Архитектура характеризуется суб-линейной вычислительной сложностью по отношению к увеличению частоты кадров, что делает ее пригодной для анализа видеопотоков в реальном времени [4].

Применение SlowFast для распознавания действий в реальном времени

В системах анализа видеоданных алгоритм SlowFast применяется для классификации действий человека в реальном времени. В качестве признакового пространства используются пространственно-временные паттерны, извлекаемые из видеопоследовательностей: изменения положения ключевых точек тела, траектории движения конечностей, взаимодействие с объектами окружения, а также контекстная информация о сцене. На основе этих данных формируются векторы признаков, которые затем классифицируются по заранее заданному набору действий (например, бег, падение, поднятие руки, использование инструмента) [5].

В системах видеонаблюдения метод используется для автоматического обнаружения нештатных ситуаций: драк, падений, проникновения в запретные зоны. Кластеризация (в смысле группировки) фрагментов видео с близкими признаками позволяет строить компактные базы событий для последующего поиска. В промышленной безопасности SlowFast применяется для контроля соблюдения техники безопасности: распознавание отсутствия каски, неправильных движений при подъеме тяжестей, нахождения в опасной зоне. Благодаря высокой частоте быстрой ветви алгоритм успевает среагировать на быстрое действие с задержкой не более 100–200 мс на современном GPU [3].

В человеко-машинном взаимодействии архитектура SlowFast используется для распознавания жестов и поз в реальном времени, что позволяет создавать бесконтактные интерфейсы для управления техникой, в том числе в условиях повышенной запыленности или влажности, где традиционные сенсоры работают ненадежно. В медицинских приложениях алгоритм помогает анализировать двигательную активность пациентов, выявлять эпизоды падений у пожилых людей или фиксировать судорожные приступы. Таким образом, применение SlowFast позволяет повысить уровень автоматизации, снизить нагрузку на операторов видеонаблюдения и создать системы раннего предупреждения чрезвычайных ситуаций [6].

Преимущества и ограничения метода

Алгоритм SlowFast обладает рядом существенных преимуществ. Во-первых, высокая точность распознавания действий на стандартных бенчмарках (Kinetics-400, AVA), достигающая 79–81% top-1 accuracy при использовании предобучения. Во-вторых, разделение на два потока позволяет эффективно использовать вычислительные ресурсы: быстрая ветвь имеет легкую архитектуру, а медленная - более глубокую, что дает лучшее соотношение точность/скорость по сравнению с однородными 3D CNN. В-третьих, латеральные соединения обеспечивают обмен информацией между ветвями, что повышает робастность к изменениям темпа выполнения действия. Метод хорошо масштабируется для длинных видеопоследовательностей за счет оконной обработки [3].

Однако алгоритм имеет ограничения. Основным является высокая вычислительная сложность по сравнению с методами, работающими на отдельных кадрах (например, 2D CNN + оптический поток). Для работы в реальном времени требуются мощные GPU (например, NVIDIA Tesla V100 или A100), что ограничивает применение на встраиваемых системах. SlowFast чувствителен к разрешению видео и частоте кадров: при сильном сжатии или низком FPS (менее 15) точность падает. Кроме того, алгоритм требует размеченных видеоданных для обучения, создание которых трудоемко. Для повышения эффективности на практике применяют методы сжатия моделей (quantization, pruning), а также используют дистилляцию знаний для создания легковесных версий [5].

Сравнительный анализ SlowFast с альтернативными архитектурами

Для объективной оценки места SlowFast среди методов распознавания действий проведем сравнение с тремя основными подходами: двухпоточная сеть (Two-Stream), 3D сверточная сеть (C3D) и Video Transformer (ViViT).

Двухпоточная сеть (Two-Stream) [7] использует два независимых канала: пространственный (обрабатывающий отдельные кадры) и временной (обрабатывающий оптический поток). Несмотря на высокую точность, она требует предварительного вычисления оптического потока, что существенно увеличивает время обработки и делает её малопригодной для реального времени. SlowFast, напротив, не нуждается в предвычислении потока, так как временная информация извлекается непосредственно из видеопоследовательности с помощью быстрой ветви.

3D сверточные сети (C3D, I3D) [8] применяют трехмерные свертки одновременно по пространству и времени. Они обеспечивают хорошее пространственно-временное моделирование, но обладают высокой параметрической сложностью. Например, I3D при использовании архитектуры ResNet-50 имеет около 25 млн параметров, тогда как SlowFast с аналогичной backbone — около 35 млн, но при этом SlowFast показывает на 3–5% более высокую точность на AVA за счет разделения потоков.

Video Transformers (ViViT, TimeSformer) [9] используют механизмы самовнимания для моделирования дальних зависимостей. Они способны превосходить SlowFast по точности на крупных датасетах (например, 84.8% на Kinetics-400 против 81.0% у SlowFast), но требуют на порядок больше вычислительных ресурсов (до 1000+ GFLOPS против 200–300 GFLOPS у SlowFast). Для систем реального времени с ограниченным бюджетом вычислений SlowFast остается более сбалансированным решением.

Таким образом, SlowFast занимает промежуточное положение между классическими двухпоточными методами и современными трансформерами, обеспечивая оптимальный компромисс между точностью, скоростью и вычислительной эффективностью для большинства прикладных задач видеонаблюдения и промышленной автоматизации.

Заключение

Алгоритм SlowFast является эффективным инструментом распознавания действий человека в реальном времени на основе видеоданных. Его применение позволяет анализировать высокоскоростные видеопотоки, выявлять сложные пространственно-временные паттерны поведения и формировать основу для построения интеллектуальных систем безопасности, управления и взаимодействия. Архитектура обеспечивает компромисс между точностью и вычислительной эффективностью, что делает ее одной из базовых моделей в задачах видеоаналитики.

Несмотря на существующие ограничения, связанные с потреблением ресурсов, SlowFast продолжает активно использоваться в практических и научных исследованиях благодаря своей универсальности и превосходству над более ранними подходами (C3D, Two-Stream). Перспективным направлением развития является интеграция SlowFast с трансформерами видео (Video Vision Transformer), применение автоматического поиска архитектур (NAS) и адаптация для работы на мобильных устройствах и пограничных вычислителях. Это позволит расширить область применения метода и обеспечить его внедрение в широкий спектр приложений, от систем «умный дом» до беспилотных летательных аппаратов.

Библиографический список

[1] Х. Н. Мохамед, А. Д. Геннадьевич, и Ч. А. Николаевич, «Алгоритм идентификации аномальных действий», *Computational nanotechnology*, т. 11, вып. 3, сс. 64–80, 2024.

[2] М. В. Алиев, Д. А. Бербенцев, В. О. Немыкин, и С. М. Алиева, «Алгоритмы распознавания объектов в видеопотоке и определение свойств их взаимного расположения», *Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки*, вып. 2 (341), сс. 27–34, 2024.

[3] C. Feichtenhofer, H. Fan, J. Malik, и K. He, «SlowFast Networks for Video Recognition», в *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, окт. 2019, сс. 6201–6210. doi: 10.1109/ICCV.2019.00630.

[4] J. Carreira и A. Zisserman, «Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset», представлено на 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, июл. 2017, сс. 4724–4733. doi: 10.1109/CVPR.2017.502.

[5] J. Jiang и Y. Zhang, «An Improved Action Recognition Network With Temporal Extraction and Feature Enhancement», *IEEE Access*, т. 10, сс. 13926–13935, 2022, doi: 10.1109/ACCESS.2022.3144035.

[6] В. Н. Денисович, «Метод выявления аномальных движений людей на видеозаписях без предварительного обучения», *Труды Московского физико-технического института*, т. 17, вып. 2 (66), сс. 5–12, 2025.

[7] C. Feichtenhofer, A. Pinz, и A. Zisserman, «Convolutional Two-Stream Network Fusion for Video Action Recognition», в *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, июн. 2016, сс. 1933–1941. doi: 10.1109/CVPR.2016.213.

[8] K. Hara, H. Kataoka, и Y. Satoh, «Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition», представлено на 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), IEEE Computer Society, окт. 2017, сс. 3154–3160. doi: 10.1109/ICCVW.2017.373.

[9] J. Yang, X. Dong, L. Liu, C. Zhang, J. Shen, и D. Yu, «Recurring the Transformer for Video Action Recognition», в *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, июн. 2022, сс. 14043–14053. doi: 10.1109/CVPR52688.2022.01367.