ОБЗОР ПОДХОДОВ К ОБНАРУЖЕНИЮ СОБЫТИЙ В ВИДЕОПОТОКЕ

Бокиев Н.С., оператор 4 научной роты ФГАУ «Военный инновационный технополис «ЭРА».

Аннотация

В статье описывается задача обнаружения событий и действий в видеопотоке, излагаются сложности и проблемы, возникающие при ее решении, а также подходы и методы, которые использовались до эпохи глубокого обучения и применяются сейчас.

Ключевые слова: компьютерное зрение, понимание видео, распознавание действий, распознавание событий, классификация видео, машинное обучение, машинное зрение, глубокое обучение

Компьютеры становятся с каждым днем все умнее и умнее. На текущий момент компьютеры уже умеют решать сложные задачи, с которыми раннее мог справиться только человек. Они автоматизируют многие интеллектуальные задачи - находят сложные взаимосвязи, закономерности в данных, понимают ауди, текст, видят объекты на изображении, описывают сцены и другие подобные задачи. В этой статье в центре внимания одна из задач компьютерного зрения – обнаружение событий в видеопотоке.

Интеллектуальная обработка видео является актуальной задачей, потому что объем видеоконтента растет с большой скоростью, камер видеонаблюдения становится все больше и больше. Человек уже не способен просмотреть весь видеоматериал, найти нужное видео или нужную сцену при работе с большим массивом данных. Так же обстоят дела с системами видеонаблюдения – в идеальном случае для обеспечения безопасности за каждой камерой необходимо закрепить человека, что не только неэффективно, но и небезопасно, так как люди устают, засыпают и ошибаются.

Люди научились хорошо находить объекты на изображении, локализовывать их, находить точные границы, распознавать лица, но несмотря на большой прогресс в развитии машинного зрения, понимание видео остается до сих пор сложной задачей из-за разнообразия действий и событий.

Особенность данной задачи в том, что для понимания необходима не только статическая информация на каждом кадре,

но и контекст видео. К сложностям распознавания событий в видеопоследовательности можно отнести:

- Большие вычислительные затраты
  Простая сверточная нейронная сеть имеет примерно 5 млн параметров, в то время как такая же трехмерная сверточная нейронная сеть имеет примерно 33 млн параметров. Для обучения таких моделей требуется больше ресурсов и времени, что замедляет проведение экспериментов.
- Захват длинного контекста
  Помимо пространственной информации, для распознавания события необходима временная информация. И при этой модель должна быть инварианта к движениям самой камеры во времени.
- Проектирование архитектуры моделей
  Особенность заключается в проектировании такого дизайна, которая учитывала бы пространственно-временную информацию. А таких вариантов множество, они нетривиальны и дороги в оценке. Среди множества архитектур можно выделить базовые три идеи:
  - одна сеть сбора пространственно-временной информации против двух отдельных
  - слияние прогнозов между несколькими клипами
  - сквозное обучение против выделения признаков и классификации видео отдельно

Обнаружение событий сводится к решению задачи классификации, поэтому можно использовать подходы анализа изображения для видео. Решая поставленную задачу с помощью классических подходов компьютерного зрения[1], алгоритм можно разделить на 3 этапа (Рисунок 1):

1. Извлечение локальных признаков.
2. Объединение локальных признаков до уровня признаков видео.
   Один из популярных методов – мешок визуальных слов, полученных с помощью иерархической или k-means кластеризации.
3. классификация, например, с помощью SVM или RF

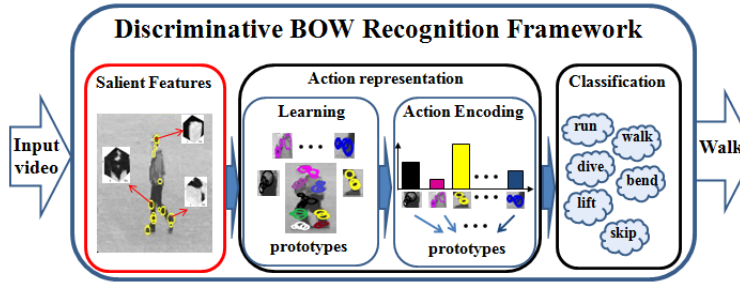

Рисунок 1 – Схема распознавания событий классическими алгоритмами компьютерного зрения.

В 2013 году появились 3D свертки, которые использовались для распознавания действий без особой помощи [2].

Затем в 2014 году вышли две прорывные исследовательские работы, основное отличие между ними — это способы объединения пространственной и временной информации.

Первой подход - однопотоковые сети[3]. Авторы статьи исследовали несколько способов объединения временной информации из последовательных кадров с использованием предварительно обученных сверток 2D.

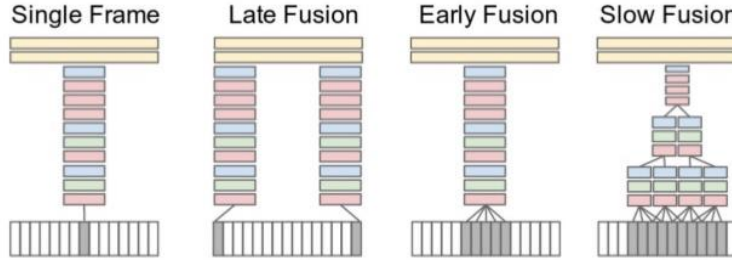

Рисунок 2 - однопотоковые сети.

На рисунке 2 показаны варианты объединения информации:

- Single frame - использует одну общую архитектуру, которая объединяет информацию из всех кадров на последнем этапе.

- Late fusion - использует две сети с общими параметрами, входные кадры которых расположены с интервалом в 15 кадров друг от друга. Предсказания объединяются в конце
- Early fusion - объединяет в первом слое путем свертки более 10 фреймов.
- Slow fusion - объединяет в несколько этапов. Это компромиссный вариант между early и late fusion. Для получения итогового результата прогнозы отобранных клипов из видео усредняются.

Несмотря на большое количество экспериментов, авторы обнаружили, что результаты были значительно хуже современных алгоритмов на основе признаков, полученных вручную.

Неудаче были приписаны следующие причины:

1. Изученные пространственно-временные признаки не захватывали особенности движения
2. Набор данных был недостаточно разнообразным, поэтому было затруднительно изучить детали событий.

Второй подход - двухпотоковые сети[4]. В этой новаторской работе авторы опираются на провалы предыдущей работы. Учитывая жесткость глубоких архитектур для изучения особенностей движения, авторы явно смоделировали элементы движения в виде суммированных векторов оптического потока. Таким образом, вместо единой сети для пространственного контекста, эта архитектура имеет две отдельные сети - одну для пространственного контекста (предварительно обученную), другую для контекста движения (Рисунок 3). Входом в пространственную сеть является один кадр видео. Авторы экспериментировали со входом во временную сеть и обнаружили, что двунаправленный оптический поток, сложенный в течение 10 последовательных кадров, показал наилучшие результаты. Два потока были обучены отдельно и объединены с использованием SVM. Окончательный прогноз был таким же, как и в предыдущей статье, то есть усреднение по выборочным кадрам.

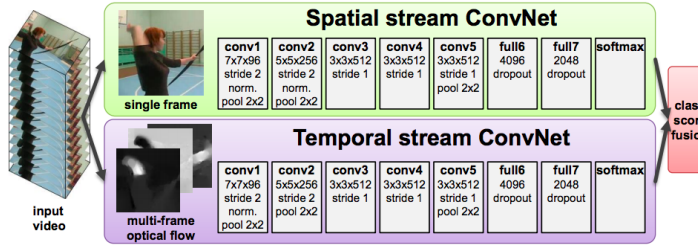

Рисунок 3 - двухпотоковые сети.

Несмотря на то, что этот метод улучшил производительность однопоточного метода за счет явного захвата локального временного движения, все же было несколько недостатков:

- Поскольку предсказания уровня видео были получены из усредненных предсказаний по выборочным клипам, продолжительная временная информация все еще отсутствовала в изученных функциях.
- Так как обучающие клипы равномерно выбираются из видео, они страдают от проблемы присвоения меток несоответствия.
- Метод включал в себя предварительную вычисление векторов оптического потока и хранение их отдельно, а также обучение для обоих потоков было раздельным, что подразумевало, что обучение на ходу все еще длинная дорога.

Это два базовых подхода, на которые основываются новые статьи. Основные идеи архитектур показаны на рисунке 4, а качество классификации конкретных реализаций этих идей показаны в таблице 1.

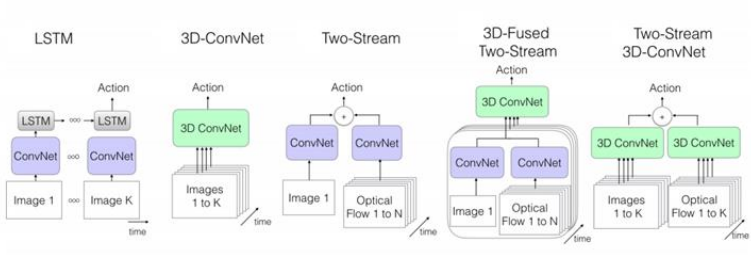

Рисунок 4 - основные идея дизайна нейронных сетей для распознавания событий

Таблица 1 – Качество классификации на наборе данных UCF101-split1

Score	Comment
94.0	TSN (RGB + Flow) [5]
94.2	TSN (RGB + Flow + Warped flow)
92.7	ActionVLAD[6]
93.6	ActionVLAD + iDT
89.8	Hidden Two Stream[7]
92.5	Hidden Two Stream + TSN
93.4	Two Stream I3D[8]
98.0	Imagenet + Kinetics pre-training
90.3	T3D[9]
91.7	T3D + Transfer
93.2	T3D + TSN
82.3	C3D (1 net) + linear SVM[10]
85.2	C3D (3 nets) + linear SVM
90.4	C3D (3 nets) + iDT + linear SVM

Интерес к задаче распознавания событий в видеопотоке не угасает. Помимо описанных подходов, есть и более сложные, например, мультимодальные многоуровневые модели, которые для классификации события используют не только последовательность кадров, но и аудиодорожки или сценария диалога. Но подобные модели сложнее на текущий момент сравнить с другими, так как открытых тестовых данных для сравнения алгоритмов пока отсутствуют.

Литература

1. Heng Wang, Action Recognition by Dense Trajectories. [Электронный ресурс] URL: https://hal.inria.fr/inria-00583818/document
2. Shuiwang Ji, 3D Convolutional Neural Networks for Human Action Recognition. [Электронный ресурс] URL: https://www.semanticscholar.org/paper/3D-Convolutional-Neural-Networks-for-Human-Action-Ji-Xu/80bfcf1be2bf1b95cc6f36d229665cdf22d76190
3. Andrej Karpathy , Large-scale Video Classification with Convolutional Neural Networks. [Электронный ресурс] URL: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/42455.pdf
4. Karen Simonyan, Two-Stream Convolutional Networks for Action Recognition in Videos. [Электронный ресурс] URL: https://arxiv.org/pdf/1406.2199.pdf
5. Jeff Donahue, Long-term Recurrent Convolutional Networks for Visual Recognition and Description. [Электронный ресурс] URL: https://arxiv.org/abs/1411.4389
6. Rohit Girdhar, ActionVLAD: Learning spatio-temporal aggregation for action classification. [Электронный ресурс] URL: https://arxiv.org/pdf/1704.02895.pdf
7. Yi Zhu, Hidden Two-Stream Convolutional Networks for Action Recognition. [Электронный ресурс] URL: https://arxiv.org/abs/1704.00389
8. Joao Carreira, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. [Электронный ресурс] URL: https://arxiv.org/abs/1705.07750
9. Ali Diba, Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification. [Электронный ресурс] URL: https://arxiv.org/abs/1711.08200
10. Du Tran, Learning Spatiotemporal Features with 3D Convolutional Networks. [Электронный ресурс] URL: https://arxiv.org/pdf/1412.0767.pdf