ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2022, № 2, с. 100–106

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

УДК 004.93

РАСПОЗНАВАНИЕ ДВИЖЕНИЙ ЧЕЛОВЕКА ПО ВИДЕОДАННЫМ

© 2022 г. М. С. Каприелова[a],*, В. Ю. Леонов[b], Р. Г. Нейчев[a],**

[a] *МФТИ, Москва, Россия*

[b] *Федеральный исследовательский центр "Информатика и управление" РАН, Москва, Россия*

**e-mail: kaprielova.ms@phystech.edu*

***e-mail: neychev@phystech.edu*

Поступила в редакцию 10.06.2021 г.
После доработки 19.10.2021 г.
Принята к публикации 29.11.2021 г.

Проанализированы различные подходы к задаче распознавания движений человека по видеоданным. Задача рассматривается в двух постановках: обучения с учителем и обучения без подготовки (zero-shot learning). Для приведенных подходов представлен анализ их свойств и указана их общность с подходами к обработке естественного языка. Отдельное внимание уделено вопросам использования дополнительных модальностей и работе в условиях недостатка размеченных данных.

DOI: 10.31857/S0002338822020093

Введение. Задача распознавания движений человека по видеоданным актуальна для множества прикладных областей. Примерами отраслей являются виртуальная реальность, восстановительная медицина и здравоохранение, безопасность и видеонаблюдение, обеспечение безопасности на производствах, спорт, робототехника, взаимодействие компьютеров и человека и др. Распознавание движений человека может помочь рассчитать взаимодействие нескольких агентов, что необходимо, например, при разработке программного обеспечения для беспилотных автомобилей. Решение задачи распознавания движений человека осложняется тем, что на то, как человек двигается, влияет множество факторов от взаимодействия с внешней средой до физических возможностей, телосложения и эмоционального состояния конкретного человека. Кроме того, движения человека могут быть обусловлены внешними воздействиями. Количество данных для тренировки сетей довольно ограниченно, а разметка датасетов под распознавание движений – достаточно трудоемкая задача. Помимо количества данных играют роль и условия, в которых они были сняты. В связи с этим большинство алгоритмов хорошо работают только на данных какого-то конкретного типа. Для большинства практических применений решений задачи распознавания движений человека по видеоданным важным критерием является возможность работы в реальном времени и на ограниченных ресурсах. Задача распознавания движений человека имеет множество разновидностей: распознавание движений всего тела, мимики, жестов и др. В статье рассматриваются задачи распознавания движений всего тела в двух постановках: обучение с учителем и решения zero-shot постановки задачи, а также формулируются ответы на три вопроса:

1) как влияет наличие дополнительных модальностей данных на точность решения задачи распознавания движений человека,

2) как подойти к задаче распознавания движений в условиях недостатка размеченных данных,

3) как могут помочь подходы из области обработки естественного языка (и, в частности, языкового моделирования) в решении задачи распознавания движений человека?

1. Постановка задачи. Сначала изучается задача обучения с учителем. Задана выборка видеозаписей, содержащих движения человека, $D = \{X, Y\}$, где X – множество элементов выборки, Y – конечное множество меток класса. Необходимо найти оптимальный классификатор $f(x)$, такой, что

$$f^* = \arg\min_f L(f, D),$$

где $L(f, D)$ – заданная функция потерь, например, количество ошибок классификации

$$L_{mis} = \sum_{(x,y)\in D} [f(x) \neq y].$$

Здесь $x \in X$, $y \in Y$, $[\cdot]$ – скобка Айверсона, функция, возвращающая 1, если аргумент является истинным утверждением и 0, если аргумент ложный.

Затем рассматривается постановка обучения без подготовки (zero-shot learning). Аналогичным образом задана выборка видеозаписей, содержащих движения человека, $D=\{X, Y\}$. При этом множество Y представляет собой объединение двух непересекающихся множеств: $Y = Y_{train} \cup Y_{test}$, где $Y_{train} \cap Y_{test} = \varnothing$, Y_{test} – метки классов тестовой выборки, Y_{train} – метки классов тренировочной выборки. Множество соответствующих меткам классов из Y_{test} объектов обозначим через X_{test}. Отложенной выборкой будем называть $D_{test} = \{X_{test}, Y_{test}\}$. Заметим, что в отложенной выборке встречаются исключительно метки классов из множества Y_{test}, не встречавшиеся в обучающей выборке. Необходимо найти классификатор $f(x)$, минимизирующий ошибку на отложенной выборке D_{test}, такой, что

$$f^* = \arg\min_f \mathrm{L}(f, D_{test}),$$

где $L(f, D)$ – заданная функция потерь, при этом $\forall (x, y) \in D_{test}$ верно, что $y \notin Y_{train}$.

2. Краткое описание подходов обучения с учителем. С развитием глубокого обучения появилось множество подходов к решению задачи распознавания движений человека по видеоданным. Изначально для решения задачи распознавания движений по видео использовались 2D-сверточные сети, после чего информация, извлеченная из обработки каждого кадра отдельно, суммировалась с целью учета упорядоченной во времени информации [1–3]. Например, в [3] были представлены TRN (temporal relation networks сети, учитывающие временные зависимости) и MTRN (multi-scale temporal relations networks сети, учитывающие временные зависимости в различных шкалах). Эти подходы не всегда достаточно эффективны для работы с длинными видеозаписями, что критично для некоторых областей применения (например, спорт).

При работе с данными часто возникает проблема наличия неинформативных признаков [4]. В частном случае задачи распознавания движений человека по видеоданным эта проблема проявляется в наличии неинформативных кадров в видеозаписях. Обработка этих данных повышает вычислительную сложность, но при этом несет крайне мало информации, необходимой для решения задачи. В связи с этим для снижения вычислительной сложности были предложены схемы семплирования, которые позволяют извлекать наиболее информативные кадры для их последующей обработки. Примером работы на эту тему являются исследования [2, 5]. В [5] утверждается, что при использовании схемы семплирования, основанной не на покадровом анализе видео, а на обработке фрагментов записи (SMART), можно выделить из видео только информативные кадры. Такой подход позволяет не только сократить вычислительную сложность, но и повысить точность.

Для более эффективного учета упорядоченности информации во времени использовались рекуррентные сети, например LSTM (долгая краткосрочная память) [6, 7], DB-LSTM (полносвязная двунаправленная долгая краткосрочная память) [8]. Так в [6] авторы предложили решение на основе LSTM с несколькими ядрами свертки и attention механизма, адаптированного под работу с ядрами разного размера. Авторы [8] сначала обрабатывали каждый шестой кадр видеозаписи сверточной нейросетью, а затем полученные данные подавали на вход двунаправленной LSTM с целью учета временной структуры. Иногда из видеоданных предварительно извлекаются ключевые точки, позволяющие определить позу человека, и с учетом этой информации решается задача распознавания движений человека [7, 9, 10]. Например, в [7] такие последовательности подавались на вход системе, состоящей из трех LSTM и семи сверточных сетей. В [9] авторы решали задачу распознавания движений на ключевых точках с помощью графовых нейросетей. Подобные архитектуры в последнее время все чаще появляются в публикациях, посвященных оценке позы человека.

При работе с видео появляется необходимость учитывать временную структуру данных. Для решения проблемы моделирования используются различные подходы. Так, в [11] для анализа данных различной природы (в частности, временных рядов, представленных в разных шкалах) предлагается применять несколько моделей. В частном случае задачи распознавания движений

человека для восстановления временной структуры и повышения точности часто используются дополнительные модальности. Примером добавления таких модальностей в задаче распознавания движений может являться звуковая дорожка видеозаписи [12–14], оценка позы человека [15] или оптический поток [13, 16, 17]. В [13] авторы отдельно обрабатывают звуковую дорожку и оптический поток, при этом информация о взаимодействии между модальностями извлекается до учета временной структуры. Авторы [12] разработали решение, состоящее из двух частей: первая часть основана на механизме дистилляции [18]: модель-учитель, которая обучена на видеоданных, и модель-ученик, которой на вход подается звуковая дорожка и первый кадр видеозаписи. Вторая часть представляет собой LSTM, которая избирательно обрабатывает кадры и аудиофрагменты. В [15] в начале оценивается поза человека по видеоданным, а затем конструируется пространственно-временной граф на последовательности поз. А в [16] предлагается предварительно извлекать из оптического потока признаки для использования в дальнейшем обучении моделей. Некоторые модальности значительно повышают вычислительную сложность, что привело к развитию отдельного направления исследований: распознавания движений человека по сжатым видеоданным. Дело в том, что в сжатых видеозаписях содержатся векторы движения, которые кодируют движения отдельных блоков пикселей [14, 19, 20]. Применение их вместо оптического потока позволяет значительно снизить вычислительную сложность, но имеет свои минусы: потеря информации о связях между различными модальностями, падение точности предсказаний в связи с потерей части информации при переходе от оптического потока к векторам движения, повышение зашумленности данных. В [20] представлены подходы к обучению, позволяющие повысить качество предсказания с помощью векторов движения. В [19] проведено сравнение нескольких стратегий обучения систем учитель-ученик, где модель-учитель обучается на оптическом потоке, а модель-ученик – на векторах движения (MV-CNN).

Для решения задачи распознавания движений человека по видеоданным применяются и 3D-сверточные сети. Пример такой модели – Two-Stream I3D [21]. Использование нейронных сетей такого типа позволяет учитывать пространственную и временную информацию одновременно благодаря применению операции свертки сразу к последовательности кадров [22, 23]. Внедрение 3D-CNN (3D-сверточные сети) хоть и позволило повысить точность решений, но сделало вычисления значительно более вычислительно затратными. В рамках борьбы с этой проблемой авторы [24–26] представили разнообразные псевдосвертки, позволяющие снизить вычислительную сложность. Одна из моделей такого типа – R[2 + 1]D [27].

В литературе также встречаются решения задачи распознавания движений человека по видеоданным, использующие информацию о связях между различными модальностями для повышения качества моделей. Методы извлечения дополнительной информации из отношений между модальностями варьируются от билинейного пулинга в общем пространстве [28] до зарекомендовавшей себя в задачах языкового моделирования архитектуры Transformer [14, 29, 30]. Так, авторы [29, 30] обучали модели на базе архитектуры Transformer одновременно на графических и текстовых данных. В [31] обучение проводилось на визуальных и аудиоданных, а в [32, 33] – на визуальных, аудио- и текстовых данных.

В последнее время в задачах компьютерного зрения набирает популярность использование механизма self-attention, стоящего в основе архитектуры Transformer. Эта тенденция не обошла стороной и распознавание движений человека по видеоданным. Так, с выходом Vision Transformer (ViT) [34] появились публикации, посвященные применению подобной архитектуры для задачи классификации видео [35, 36]. В исследовании [36] было экспериментально установлено, что временной и пространственный attention механизмы, примененные отдельно к каждому блоку, значительно повышают качество классификации. В [14] успешно использовали трансформер для распознавания движений человека на сжатых видеоданных, принимая на вход изображения, данные о движении и звуковую дорожку. Работа интересна тем, что предложенное в ней решение позволяет учитывать информацию о связях между различными модальностями.

В табл. 1 приведены результаты, достигнутые с помощью различных архитектур на датасете UCF101 [37], который представляет собой 13320 видеозаписей, на каждой из которых представлено одно действие, выполняемое человеком. Средняя длительность видео составляет 180 кадров, всего содержится 101 действие.

В [38] представлена модель (VidTr-L) на основе Vision Transformer, в [39] использовали оптический поток в качестве модальности (Optical Flow Guided Feature), авторы [40] выбрали LSTM (TS-LSTM), а в основе [41] лежат графовые сети (MLGCN).

3. Методы обучения без подготовки (zero-shot обучение). Часто область практического применения распознавания движений человека по видеоданным предполагает возможность работы с

Таблица 1. Результаты, достигнутые с использованием различных архитектур на датасете UCF101 [37]

Модель	Статья	Точность на топ-3	Год выпуска	Особенности архитектуры
SMART	[5]	98.64	2020	Использование стратегии выделения наиболее информативных кадров
Two-Stream I3D (imagenet + kinetics pretraining)	[21]	97.8	2017	3D-сверточные сети
R[2 + 1]D – TwoStream (kinetics pretrained)	[27]	97.3	2017	2D + псевдосвертки
VidTr-L	[38]	96.7	2021	Трансформер
Optical flow Guided feature	[39]	96	2017	Использование оптического потока
TS-LSTM	[40]	94.1	2017	LSTM
MV-CNN	[19]	86.4	2016	Учитель (на оптическом потоке), ученик (на векторах движения)
MLGCN	[41]	63.27	2019	Графовые сети

движениями, которых не было в обучающей выборке. Это обусловлено тем, что сбор достаточного количества размеченных данных для моделей – дорогая и трудозатратная задача. В связи с этим активно исследуется область zero-shot обучения для задачи распознавания движений. Zero-shot обучение нацелено на возможность обобщения модели для распознавания движений на виды движений, которые не были представлены в тренировочной выборке. Для распознавания незнакомых движений такой генерализованной модели не понадобится размеченных данных.

Для каждого класса (вида детектируемого движения) строятся векторные представления в семантическом пространстве. Таким образом удается построить связи между классами из тренировочной и тестовой выборок. Чаще всего используется метод поиска ближайшего соседа. Существует несколько подходов к построению векторных представлений для решения задачи zero-shot обучения. Каждый из этих подходов имеет свои плюсы и минусы. В [42, 43] для построения представлений атрибуты действий задаются вручную. Такой метод не всегда эффективен, так как действиям иногда непросто определить корректные атрибуты.

Использование визуальных (основанных на графических данных) векторных представлений интуитивно, но не всегда удобно, так как иногда данных недостаточно для построения информативных представлений. Способы получения информативных векторных представлений из графических данных исследовались в [44–46]. В [44] для построений векторных представлений используются объекты, распознанные на видео нейронной сетью (O2A). Авторы [45] связывают не только атрибуты и информацию о распознанных объектах из тестовой выборки с атрибутами и информацией о распознанных объектах из тренировочной выборки соответственно, но и атрибуты и информацию о распознанных объектах попарно (TS-GCN). С одной стороны, такой подход имеет преимущества: объектам легко сопоставить векторные представления, которые хорошо обобщаются. С другой стороны, такой подход к построению представлений не учитывает взаимодействия объектов. В некоторых исследованиях для создания векторных представлений используются 3D-нейронные сети, способные выделять пространственную и временную информацию из видео, что является несомненным преимуществом. В [47] предложен иной подход: авторы предлагают применять модель похожей архитектуры для предсказания векторных представлений для названий действий (E2E). Минусом этого подхода является склонность к переобучению на тренировочной выборке. Альтернативный способ построения векторных представлений – использование в качестве данных для построения названий действий. Например, в [48] предлагается решение, основанное на помехоустойчивом кодировании (ZSECOC). Но и этот, казалось бы, интуитивно понятный и эффективный подход не универсален. Слова могут иметь разное значение в зависимости от контекста, а названия действий могут содержать слова, употребленные в переносном смысле.

В последнее время популярность набирает подход, объединяющий в себе векторные представления, основанные на тексте и визуальных данных. В качестве примера может выступить работа [49]. В ней обрабатывают описания действий для получения представлений из текста и соответствующие действиям изображения (ASR). Любопытно, что сами видеоданные для

Таблица 2. Результаты, достигнутые с использованием различных подходов на датасете UCF101 [37]

Модель	Статья	Точность на топ-1	Год выпуска
ER-ZSAR	[50]	51.8	2021
E2E	[47]	48	2020
TS-GCN	[45]	34.2	2019
O2A	[44]	30.3	2015
ASR	[49]	24.4	2017
ZSECOC	[48]	15.1	2017

построения векторных представлений не используются. Авторы [50] обрабатывали предложения, которые описывают действия, предобученным BERT [51] и линейным преобразованием (ER-ZSAR). Таким образом они получили информативные векторные представления. Также строятся представления на основании пространственно-временной информации с помощью 3D-нейросети из видеоданных и применяется информация об объектах, распознанных на видео.

В табл. 2 приведены результаты, достигнутые с использованием различных подходов на датасете UCF101 [37].

Отдельно рассматривается вопрос чистоты экспериментов в области zero-shot обучения. Часто в обучающую выборку попадают объекты из тестовой выборки. В частности, на исследование этой проблемы направлено внимание в [52], где предложен новый протокол составления обучающей и тестовой выборок, позволяющий оценивать работу решений zero-shot обучения на тестовых выборках, которые не пересекаются с тренировочными.

Заключение. Использование дополнительных модальностей позволяет решать задачу распознавания движений с более высокой точностью/качеством как в случае обучения с учителем, так и в постановке типа zero-shot. В течение последних лет основные подходы перешли от работы исключительно с визуальным доменом к использованию как визуальных данных, так и данных другого рода.

3D-свертки и трансформеры применяются для выявления связей/зависимостей на “низком уровне” не только для пикселей, но и для данных, упорядоченных во времени.

Языковые модели вдохновляют развитие других областей. Предложенные для задач языкового моделирования архитектуры достаточно универсальны. Это указывает на их способность не только улавливать разноплановые низкоуровневые зависимости в языке (на уровне символов/морфем/слов), но и в других данных сложной структуры, обусловленных на какой-то достаточно строго формализованный процесс (например, движения человека).

СПИСОК ЛИТЕРАТУРЫ

1. *Karpathy A., Toderici G., Shetty S., Leung T., Sukthankar R., Fei-Fei L.* Large-scale Video Classification with Convolutional Neural Networks // Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Columbus, 2014. С. 1725–1732.
2. *Wang L., Xiong Y., Wang Z., Qiao Y., Lin D., Tang X., Van Gool L.* Temporal Segment Networks: Towards Good Practices for Deep Action Recognition // Europ. Conf. on Computer Vision. Cham: Springer, 2016. С. 20–36.
3. *Zhou B., Andonian A., Oliva A., Torralba A.* Temporal Relational Reasoning in Videos // Proc. Europ. Conf. on Computer Vision (ECCV). Munich, 2018. С. 803–818.
4. *Neichev R.G., Katrutsa A.M., Strizhov V.V.* Robust Selection of Multicollinear Features in Forecasting // Industrial Laboratory. Diagnostics of Materials. 2016. Т. 82. № 3. С. 68–74.
5. *Gowda S. N., Rohrbach M., Sevilla-Lara L.* SMART Frame Selection for Action Recognition // arXiv preprint arXiv:2012.10671. 2020.
6. *Agethen S., Hsu W.H.* Deep Multi-kernel Convolutional LSTM Networks and an Attention-based Mechanism for Videos // IEEE Transactions on Multimedia. 2019. Т. 22. № 3. С. 819–829.
7. *Li C., Wang P., Wang S., Hou Y., Li W.* Skeleton-based Action Recognition Using LSTM and CNN // IEEE Intern. Conf. on Multimedia & Expo Workshops (ICMEW). Hong Kong: IEEE, 2017. С. 585–590.
8. *Ullah A., Ahmad J., Muhammad K., Sajjad M., Baik S.W.* Action Recognition in Video Sequences Using Deep Bi-directional LSTM with CNN Features // IEEE Access. 2017. Т. 6. С. 1155–1166.
9. *Li S., Yi J., Farha Y. A., Gall J.* Pose Refinement Graph Convolutional Network for Skeleton-Based Action Recognition // IEEE Robotics and Automation Letters. 2021. Т. 6. № 2. С. 1028–1035.

10. *Peng W., Shi J., Xia Z., Zhao G.* Mix Dimension in Poincaré Geometry for 3d Skeleton-based Action Recognition // Proc. of the 28th ACM Intern. Conf. on Multimedia. Seattle, 2020. С. 1432–1440.
11. *Neichev R.G.* Multimodel Forecasting of Multiscale Time Series in Internet of Things // Proc. 11th Intern. Conf. on Intelligent Data Processing: Theory and Applications, Barcelona, Spain, 2016.
12. *Gao R., Oh T.H., Grauman K., Torresani L.* Listen to Look: Action Recognition by Previewing Audio // Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle, 2020. С. 10457–10467.
13. *Kazakos E., Nagrani A., Zisserman A., Damen D.* Epic-fusion: Audio-visual Temporal Binding for Egocentric Action Recognition // Proc. IEEE/CVF Intern. Conf. on Computer Vision. Seoul, 2019. С. 5492–5501.
14. *Chen J., Ho C.M.* MM-ViT: Multi-Modal Video Transformer for Compressed Video Action Recognition // arXiv preprint arXiv:2108.09322. 2021.
15. *Yan S., Xiong Y., Lin D.* Spatial Temporal Graph Convolutional Networks for Skeleton-based Action Recognition // Thirty-second AAAI Conf. on Artificial Intelligence. 2018. New Orleans.
16. *Kwon H., Kim M., Kwak S., Cho M.* Motionsqueeze: Neural Motion Feature Learning for Video Understanding // Europ. Conf. on Computer Vision. Cham: Springer, 2020. С. 345–362.
17. *Simonyan K., Zisserman A.* Two-stream Convolutional Networks for Action Recognition in Videos // arXiv preprint arXiv:1406.2199. 2014.
18. *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // arXiv preprint arXiv:1503.02531. 2015.
19. *Zhang B., Wang L., Wang Z., Qiao Y., Wang H.* Real-time Action Recognition with Enhanced Motion Vector CNNs // Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas. 2016. С. 2718–2726.
20. *Zhang B., Wang L., Wang Z., Qiao Y., Wang H.* Real-time Action Recognition with Deeply Transferred Motion Vector CNNs // IEEE Transactions on Image Processing. 2018. Т. 27. № 5. С. 2326–2339.
21. *Carreira J., Zisserman A.* Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset // Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu, 2017. С. 6299–6308.
22. *Ji S., Xu W., Yang M., Yu K.* 3D Convolutional Neural Networks for Human Action Recognition // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012. Т. 35. № 1. С. 221–231.
23. *Tran D., Bourdev L., Fergus R., Torresani L., Paluri M.* Learning Spatiotemporal Features with 3D Convolutional Networks // Proc. IEEE Intern. Conf. on Computer Vision. Santiago, 2015. С. 4489–4497.
24. *Chen J., Hsiao J., Ho C.M.* Residual Frames with Efficient Pseudo-3D CNN for Human Action Recognition // arXiv preprint arXiv:2008.01057. 2020.
25. *Qiu Z., Yao T., Mei T.* Learning Spatio-temporal Representation with Pseudo-3d Residual Networks // Proc. IEEE Intern. Conf. on Computer Vision. Venice. 2017. С. 5533–5541.
26. *Tran D., Wang H., Torresani L., Feiszli M.* Video Classification with Channel-separated Convolutional Networks // Proc. IEEE/CVF Intern. Conf. on Computer Vision. Seoul. 2019. С. 5552–5561.
27. *Tran D., Wang H., Torresani L., Ray J., LeCun Y., Paluri M.* A Closer Look at Spatiotemporal Convolutions for Action Recognition // Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City. 2018. С. 6450–6459.
28. *Fukui A., Park D.H., Yang D., Rohrbach A., Darrell T., Rohrbach M.* Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding // arXiv preprint arXiv:1606.01847. 2016.
29. *Radford A., Kim J.W., Hallacy C. et al.* Learning Transferable Visual Models from Natural Language Supervision // arXiv preprint arXiv:2103.00020. 2021.
30. *Luo H., Ji L., Zhong M. et al.* Clip4clip: An Empirical Study of Clip for End to End Video Clip Retrieval // arXiv preprint arXiv:2104.08860. 2021.
31. *Lee S., Yu Y., Kim G. et al.* Parameter Efficient Multimodal Transformers for Video Representation Learning // arXiv preprint arXiv:2012.04124. 2020.
32. *Tsai Y.H.H., Bai S., Liang P.P. et al.* Multimodal Transformer for Unaligned Multimodal Language Sequences // Proc. of the Conf. Association for Computational Linguistics. Meeting: NIH Public Access. 2019. Т. 2019. С. 6558.
33. *Zadeh A., Mao C., Shi K. et al.* Factorized Multimodal Transformer for Multimodal Sequential Learning // arXiv preprint arXiv:1911.09826. 2019.
34. *Dosovitskiy A., Beyer L., Kolesnikov A. et al.* An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale // arXiv preprint arXiv:2010.11929. 2020.
35. *Arnab A., Dehghani M., Heigold G. et al.* Vivit: A Video Vision Transformer // arXiv preprint arXiv:2103.15691. 2021.
36. *Bertasius G., Wang H., Torresani L.* Is Space-Time Attention All You Need for Video Understanding? // arXiv preprint arXiv:2102.05095. 2021.
37. *Soomro K., Zamir A. R., Shah M.* A Dataset of 101 Human Action Classes from Videos in the Wild // Center for Research in Computer Vision. 2012. Т. 2. № 11.

38. *Li X., Liu C., Zhang Y. et al.* VidTr: Video Transformer Without Convolutions // arXiv preprint arXiv:2104.11746. 2021.
39. *Sun S., Kuang Z., Sheng L. et al.* Optical Flow Guided Feature: A Fast and Robust Motion Representation for Video Action Recognition // Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Salt Lake City, 2018. С. 1390–1399.
40. *Ma C.Y., Chen M.H., Kira Z. et al.* TS-LSTM and Temporal-inception: Exploiting Spatiotemporal Dynamics for Activity Recognition // Signal Processing: Image Communication. 2019. Т. 71. С. 76–87.
41. *Mazari A., Sahbi H.* MLGCN: Multi-Laplacian Graph Convolutional Networks for Human Action Recognition // BMVC. Cardiff, 2019. С. 281.
42. *Liu J., Kuipers B., Savarese S.* Recognizing Human Actions by Attributes // CVPR. IEEE Colorado Springs, 2011. С. 3337–3344.
43. *Zellers R., Choi Y.* Zero-shot Activity Recognition with Verb Attribute Induction // arXiv preprint arXiv:1707.09468. 2017.
44. *Jain M., Van Gemert J.C., Mensink T. et al.* Objects2Action: Classifying and Localizing Actions Without any Video Example // Proc. IEEE Intern. Conf. on Computer Vision. Santiago. 2015. С. 4588–4596.
45. *Gao J., Zhang T., Xu C.* I Know the Relationships: Zero-shot Action Recognition via Two-stream Graph Convolutional Networks and Knowledge Graphs // Proc. AAAI Conf. on Artificial Intelligence. Honolulu, 2019. Т. 33. № 01. С. 8303–8311.
46. *Gan C., Lin M., Yang Y. et al.* Concepts not Alone: Exploring Pairwise Relationships for Zero-shot Video Activity Recognition // Thirtieth AAAI Conf. on Artificial Intelligence. Phoenix, 2016.
47. *Brattoli B., Tighe J., Zhdanov F. et al.* Rethinking Zero-shot Video Classification: End-to-end Training for Realistic Applications // Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle, 2020. С. 4613–4623.
48. *Qin J., Liu L., Shao L. et al.* Zero-shot Action Recognition with Error-correcting Output Codes // Proc. IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu, 2017. С. 2833–2842.
49. *Wang Q., Chen K.* Alternative Semantic Representations for Zero-shot Human Action Recognition // Joint Europ. Conf. on Machine Learning and Knowledge Discovery in Databases. Cham: Springer, 2017. С. 87–102.
50. *Chen S., Huang D.* Elaborative Rehearsal for Zero-shot Action Recognition // arXiv preprint arXiv:2108.02833. 2021.
51. *Devlin J., Chang M.W., Lee K., Toutanova K.* Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding // arXiv preprint arXiv:1810.04805. 2018.
52. *Gowda S.N., Sevilla-Lara L., Kim K., Keller F., Rohrbach M.* A New Split for Evaluating True Zero-Shot Action Recognition // arXiv preprint arXiv:2107.13029. 2021.