УДК 004.93 **Информационные технологии**

Азаренко Константин Алекандрович, *студент кафедры ИУ4-КФ «Программное обеспечение, информационные технологии», КФ МГТУ им. Н.Э. Баумана*

Каунг Мьят Ньейн, *студент кафедры ИУ4-КФ «Программное обеспечение, информационные технологии», КФ МГТУ им. Н.Э. Баумана*

Белов Юрий Сергеевич, *кандидат физико-математических наук, доцент кафедры ИУ4-КФ «Программное обеспечение, информационные технологии», КФ МГТУ им. Н.Э. Баумана*

ОБЗОР МЕТОДОВ ДЛЯ РАСПОЗНАВАНИЯ ДЕЙСТВИЙ ЧЕЛОВЕКА

Аннотация: В данной статье рассматриваются современные подходы распознавания действий человека. Также будут описаны особенности использования, преимущества и недостатки данных подходов. В современном мире благодаря развитию сенсорных и визуальных технологий, системы распознавания действий человека стали популярны. Существует два основных подхода. Первый и традиционный подход основан на дескрипторах. Этот подход наиболее популярен и добился замечательных результатов при распознавании общедоступных наборов данных различной сложности. Среди методов с использованием дескрипторов выделяется несколько отдельных: пространственно-временной метод, метод локального бинарного шаблона и метод с использованием нечеткой логики. Второй подход основан на использовании нейронных сетей. Методы с использованием нейронных сетей делятся на два типа моделей: с контролируемым и неконтролируемым обучением. Эти методы пришли на замену методам с использованием дескрипторов, так как могут работать с необработанными данными и автоматизируют процесс классификации, а в случае дескрипторов методы имеют достаточно много ограничений. Самыми популярными моделями

являются глубокая сеть доверия для неконтролируемого обучения и свёрточная нейронная сеть для контролируемого обучения.

Ключевые слова: Распознавание действий человека. Глубокое обучение. Дескрипторы контрольных точек. Локальные бинарные шаблоны.

Abstract: This article discusses modern approaches to the recognition of human actions. Features of use, advantages and disadvantages of these approaches will also be described. In the modern world, thanks to the development of sensory and visual technologies, human action recognition systems have become popular. There are two main approaches. The first and traditional approach is based on descriptors. This approach is the most popular and has achieved remarkable results in recognizing publicly available data sets of varying complexity. Among the methods using descriptors, several separate ones are distinguished: the space-time method, the local binary template method and the method using fuzzy logic. The second approach is based on the use of neural networks. Methods using neural networks are divided into two types of models: with controlled and uncontrolled learning. These methods came to replace methods using descriptors, since they can work with raw data and automate the classification process, and in the case of descriptors, the methods have quite a few limitations. The most popular models are a deep network of trust for unsupervised learning and a convolutional neural network for supervised learning.

Keywords: Human action recognition. Deep learning. Handcrafted approach. Space-Time-Based Approaches. Local Binary Pattern.

Введение. Распознавание действий человека (HAR) играет важную роль во многих реальных приложениях. Целью HAR является распознавание действий человека или группы людей, полученной с сенсоров или видеоданных, включая информацию о контексте, в котором происходит деятельность. Благодаря развитию сенсорных и визуальных технологий, системы на базе HAR широко использовались во многих реальных приложениях. В частности, распространение сенсоров малого размера

позволило интеллектуальным устройствам распознавать деятельность человека с учетом контекста. Основываясь на методологиях разработки и процессе сбора данных, подходы подразделяются на визуальные, не визуальные и смешанные. Основное различие между визуальными и другими типами сенсоров — это способ восприятия данных. Визуальные сенсоры предоставляют данные в виде 2D или 3D изображений или видео, тогда как другие сенсоры в виде одномерного сигнала [1].

Было опубликовано большое количество обзоров и обзоров по HAR и связанным с ними процессам. Однако из-за большого количества работ, опубликованных по этому вопросу, опубликованные обзоры быстро устаревают. По этой же причине написание обзорной статьи является актуальной задачей. В статье будут рассматриваться современные методы распознавания человеческой деятельности, основанные на дескрипторах и нейронных сетях.

Традиционный подход. Данный подход к распознаванию действий человека основан на дескрипторах [2]. Этот подход популярен среди сообщества HAR и добился замечательных результатов в различных общедоступных наборах данных. В этом подходе извлекаются важные функции из последовательности кадров изображений, а дескриптор функции создается с использованием специализированных детекторов функций. Затем выполняется классификация путем обучения универсального классификатора опорных векторов (SVM). Этот подход включает методы, основанные на моделях внешнего вида и пространственно-временных моделях, методы с применением локального двоичного шаблона и нечеткой логики, как показано на рисунке 1.

Рис. 1. Традиционный подход к представлению и распознаванию действий.

Пространственно-временные модели. Пространственно-временные модели имеют четыре основных компонента: детектор пространственно-временной точки (STIP) [3], дескриптор функции, конструктор словарей и классификатор. Детекторы STIP далее подразделяются на плотные и разреженные детекторы. Плотные детекторы, такие как V-FAST, детектор Гессиана, плотная выборка, плотно покрывают весь видеоконтент для обнаружения контрольных точек, в то время как разреженные детекторы, такие как кубический детектор Harris3D, и пространственно-временная модель неявной формы (STISM) [4], используют разреженное (локальное) подмножество этого содержимого. Различные детекторы STIP были разработаны различными исследователями. Дескрипторы функций также делятся на локальные и глобальные дескрипторы. Локальные дескрипторы, такие как кубический дескриптор, расширенные ускоренные надежные

функции (ESURF) и N-jet, основаны на локальной информации, такой как текстура, цвет и положение, в то время как глобальные дескрипторы используют глобальную информацию, такую как изменения освещения, фазовые изменения, и изменение скорости в видео. Создатели словаря или агрегирующие методы основаны на сумме слов (BOW) или модели состояния пространства. Наконец, для классификации используется контролируемый или неконтролируемый классификатор, как показано на рисунке 2.

Локально двоичный шаблон (LBP). Локальный двоичный шаблон (LBP) — это тип визуального дескриптора для классификации текстур. С момента своего создания было предложено несколько модифицированных версий этого описания для различных задач, связанных с классификацией, в компьютерном видении. Был предложен метод распознавания человеческого действия на основе LBP в сочетании с инвариантностью внешнего вида и методом сопоставления патчей [5]. Этот метод был протестирован на разных публичных наборах данных и оказался эффективным для распознавания действий. Другой метод распознавания активности был предложен с использованием дескриптора LBP-TOP. В этом методе объем действия был

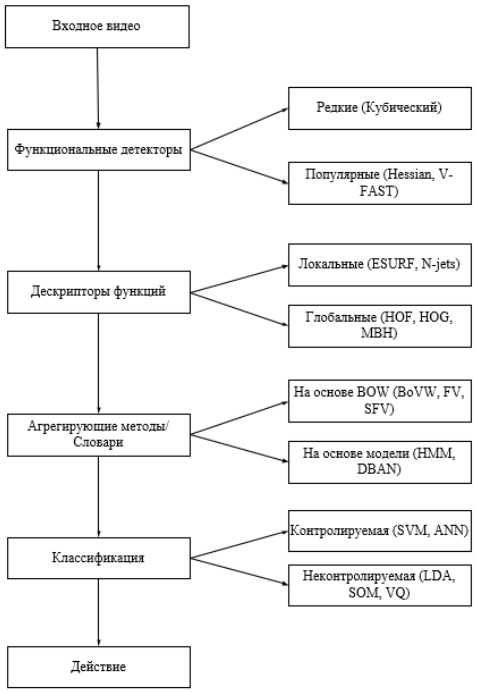

Рис. 2. Компоненты пространственно-временных подходов.

разделен на подтомы, а гистограмма функций была создана путем объединения гистограмм подтомов. Используя это представление, они закодировали движение на трех разных уровнях: уровень пикселей (одиночный бит в гистограмме), региональный уровень (гистограмма субтомника) и глобальный уровень (конкатенация суб-томов-гистограмм). Методы, основанные на LBP, также использовались для многопользовательского распознавания человеческого действия. Было предложено многопользовательский метод распознавания человеческого действия, основанный на контурных позициях и равномерном вращении-инвариантном LBP, за которым следует SVM для классификации. Недавно был введен еще один дескриптор движения с именем Motion Binary Pattern (MBP) для распознавания множественных действий. Этот дескриптор представляет собой комбинацию Volume Local Binary Pattern (VLBP) и оптического потока [6]. Этот метод был оценен на основе мультивариантного набора данных INSIA Xmas Motion Acquisition Sequences (IXMAS) и достиг 80,55% результатов распознавания.

Нечеткий логический подход. Традиционные методы распознавания человеческого действия на основе видения используют пространственные или временные признаки, за которыми следует общий классификатор для представления действий и классификации. Тем не менее, трудно масштабировать эти методы для обработки неопределенности и сложности, связанных с приложениями реального мира. Для решения этих трудностей методы, основанные на нечеткой основе, считаются лучшим выбором. Предложена нечеткая структура для распознавания человеческого действия на основе нечетких лог-полярных гистограмм и временного самоподобия для представления действия, за которым следует SVM для классификации действий [7]. Оценка предложенного метода на двух публичных наборах данных подтвердила высокую точность и пригодность для приложений реального мира. Был предложен другой метод, основанный на нечеткой логике, в котором использовались срезы силуэта и характеристики скорости движения в качестве

входных данных для нечеткой системы, и использовали метод кластеризации с нечетким средством для получения функции членства для предлагаемой системы. Результаты подтвердили лучшую точность предлагаемой нечеткой системы по сравнению с не-нечеткими системами для одного и того же публичного набора данных. Большинство методов распознавания человеческого действия зависят от вида и могут распознавать действие с фиксированного вида. Однако метод распознавания человеческих действий в режиме реального времени должен быть способен распознавать действие с любой точки зрения. Для достижения этой цели многие современные методы используют данные, захваченные несколькими камерами. Однако это не практическое решение, поскольку калибровка нескольких камер в реальных сценариях довольно сложная. Использование единственной камеры должно быть окончательным решением для распознавания инвариантного действия вида. В этих строках был предложен метод на основе нечеткой логики для распознавания инвариантного действия вида с использованием одной камеры. Этот метод извлекал человеческие контуры из нечеткой качественной модели Пуассона для оценки зрения, за которой следуют алгоритмы кластеризации для классификации вида, как показано на рисунке 6. Результаты показывают, что предложенный метод достаточно эффективен для просмотра независимого распознавания действий. Некоторые методы, основанные на нейро-нечетких системах (NFS), также были предложены для распознавания жестов и действия человека [8]. В дополнение к этому, эволюционирующие системы также очень успешны в распознавании поведения.

Методы с применением глубокого обучения. Недавние исследования показывают, что для всех наборов данных нет универсально лучших дескрипторов функций ручной работы, поэтому изучение функций непосредственно из необработанных данных может быть более выгодным. Глубокое обучение является важной областью машинного обучения, которая направлена на изучение множества уровней представления и абстракции, которые могут осмыслить такие данные, как речь, изображения и текст.

Методы, основанные на глубоком обучении, обладают способностью обрабатывать изображения/видео в их необработанных формах и автоматизировать процесс извлечения, представления и классификации объектов. В этих методах используются обучаемые экстракторы признаков и вычислительные модели с несколькими уровнями обработки для представления и распознавания действий. Основываясь на исследовании, посвященном глубокому изучению, мы классифицировали модели глубокого обучения по трем категориям:

(1) генеративные / неконтролируемые модели (например, Deep Belief Networks (DBNs), машины Deep Boltzmann (DBM), Restricted Boltzmann Машины (RBM) и регуляризованные автокодеры);

(2) Дискриминационные / контролируемые модели (например, глубокие нейронные сети (DNN), повторяющиеся нейронные сети (RNN) и сверточные нейронные сети (CNN));

(3) гибридные модели, эти модели используют характеристики обеих моделей, например Цели дискриминации могут быть полезны для результатов генеративной модели. Однако эти модели не рассматриваются здесь отдельно.

Методы с применением неконтролируемого глубокого обучения. Неконтролируемые модели глубокого обучения не требуют ярлыков целевого класса во время учебного процесса. Эти модели особенно полезны, когда помеченные данные относительно скудны или недоступны. Модели глубокого обучения исследовались с 1960-х годов, но исследователи уделяли мало внимания этим моделям. Это было связано главным образом с успешностью неглубоких моделей, таких как SVM, и отсутствием огромного объема данных, необходимых для обучения глубоким моделям.

Замечательный всплеск истории глубоких моделей был вызван работой Хинтона и др., где был внедрен высокоэффективный DBN и алгоритм обучения, за которым следует методика сокращения признаков. DBN был обучен поэтапно с использованием RBM, параметры, полученные во время этой неконтролируемой фазы предварительного обучения, были точно настроены

контролируемым образом с использованием обратного распространения. С момента внедрения этой эффективной модели большой интерес представляет применение моделей глубокого обучения для различных приложений, таких как распознавание речи, классификация изображений, распознавание объектов и распознавание человеческих действий.

Для распознавания действий был предложен метод, использующий неконтролируемое распознавание признаков из видеоданных. Авторы использовали алгоритм независимого подпространственного анализа, чтобы изучить пространственно-временные функции, сочетающие их с методами глубокого обучения, такими как сверточное и разбиение на представление действий и распознавание. Для распознавания действий человека использовались сети глубоких убеждений (ДБН), прошедшие обучение с использованием УОКР. Предлагаемый метод превзошел методы с дескрипторами. Постоянное изучение потокового видео без каких-либо ярлыков является важной, но сложной задачей. Этот вопрос был решен с использованием неконтролируемой модели глубокого обучения. Большинство наборов данных действий были записаны в контролируемой среде; распознавание действий из неограниченных видеороликов является сложной задачей.

Неконтролируемое обучение сыграло ключевую роль в возрождении интересов исследователей в глубоком обучении. Тем не менее, это было омрачено чисто контролируемым обучением, так как основной прорыв в глубоком обучении использовал CNN для распознавания объектов. Обратно, важное исследование, проведенное пионерами последних моделей глубокого обучения, предполагает, что неконтролируемое обучение будет намного больше важнее, чем его контролируемый коллега, в конечном счете, так как мы открываем мир, наблюдая его, а рассказывая имя каждого объекта. Обучение людей и животных в основном не контролируется [9].

Методы с применением контролируемого глубокого обучения. Для распознавания действий человека, наиболее часто используемой моделью с

контролируемым обучением является сверточная нейронная сеть (CNN). CNN — это модель глубокого обучения, которая показала отличную производительность при таких задачах, как распознавание образов, классификация по рукописным цифрам, классификация изображений и распознавание человеческих действий. Это иерархическая модель обучения с несколькими скрытыми слоями для преобразования входного объема в категории вывода. Его архитектура состоит из трех основных типов слоев: сверточного слоя, слоя объединения и полностью подключенного слоя. Понимание работы различных слоев CNN требует отображения этих действий в пространство пикселей, это делается с помощью Deconvolutional Networks (Deconvnets). Deconvnets используют тот же процесс, что и CNN, но в обратном порядке для отображения из пространственного пространства в пространство пикселей. Первоначально глубокий CNN использовался для представления и распознавания объектов из неподвижных изображений. Это было расширено для распознавания действий из видео с использованием штабелированных видеокадров в качестве входных данных в сеть, но результаты были хуже, чем даже мелкие представления ручной работы. Для решения этого вопроса была придумана двухпотоковая сверточная нейронная сеть для распознавания действий. Оба этих потока были реализованы как Convnet, пространственный поток распознает действие из неподвижных видеокадров, а временный поток выполняет распознавание действия из движения в виде плотного оптического потока. Впоследствии эти два потока объединяли с использованием позднего синтеза для распознавания действия. Этот метод обеспечил превосходные результаты для одного из лучших методов представления мелкой ручной работы. Тем не менее, архитектура с двумя потоками может оказаться непригодной для приложений реального времени из-за сложности вычислений [9].

Заключение. Глубокое обучение стало очень популярным направлением в рамках машинного обучения, которое превзошло традиционные подходы во многих приложениях компьютерного зрения. Очень выгодным свойством

алгоритмов глубокого обучения является их способность изучать функции из необработанных данных, что устраняет необходимость в дескрипторах и дескрипторах ручной работы. Существуют две категории моделей глубокого обучения, т. е. неконтролируемые и контролируемые модели. DBN - популярная неконтролируемая модель, которая используется для распознавания человеческих действий. Эта модель уже достигла высокой производительности в сложных наборах, данных по сравнению с традиционными коллекторами ручной работы. С другой стороны, CNN является одной из самых популярных моделей глубокого обучения. Большинство существующих обучающих представлений либо непосредственно применяют CNN к видеокадрам, либо вариации CNN для извлечения и представления пространственно-временных характеристик. Эти модели также достигли отличных результатов по сложным наборам данных о распознавании человеческой активности. До сих пор контролируемые модели глубокого обучения добились более высокой производительности, но некоторые исследования показывают, что в долгосрочной перспективе обучение без надзора будет гораздо более важным. Поскольку мир обнаружен, наблюдая за тем, что ему говорят, что имя каждого объекта, обучение людей и животных в основном не контролируется. В некоторых исследованиях сообщалось, что включение функций ручной работы в модель CNN может повысить эффективность распознавания действий.

Библиографический список:

1. Xia L., Gori I., Aggarwal J. K., Ryoo M. S. Robot-centric activity recognition from first-person rgb-d videos // Applications of Computer Vision (WACV), IEEE Winter Conference on, 2015, pp. 357–364.

2. Soomro K., Zamir A. R. Action recognition in realistic sports videos // Springer, 2014, pp. 181–208.

3. Kihl O., Picard D., Gosselin P.-H. Local polynomial space–time descriptors for action classification // Machine Vision and Applications, 2016, vol. 27, no. 3, pp. 351–361.

4. Wang H., Schmid C. Action recognition with improved trajectories // Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3551–3558.

5. Bux A., Angelov P., Habib Z. Vision based human activity recognition: a review // Springer, 2017, pp. 341–371.

6. Sargano A. B., Angelov P., Habib Z. A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition // Applied Sciences, 2017, vol. 7, no. 1, p. 110,

7. Baumann F., Ehlers A., Rosenhahn B., Liao J. Recognizing human actions using novel space-time volume binary patterns // Neurocomputing, 2016, vol. 173, pp. 54–63.

8. Ranasinghe S., Al Machot F., Mayr H. C., A review on applications of activity recognition systems with regard to performance and evaluation // International Journal of Distributed Sensor Networks, 2016, URL: https://journals.sagepub.com/doi/pdf/10.1177/1550147716665520 (дата обращения: 28.03.2019).

9. Sargano A. B., Wang X., Angelov P., Habib Z. Human action recognition using transfer learning with deep representations // Neural Networks (IJCNN), IEEE International Joint Conference, 2017, pp. 463–469.