A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets

Vijeta Sharma, Manjari Gupta, Anil Kumar Pandey, Deepti Mishra & Ajai Kumar

Taylor & Francis
Taylor & Francis Group

A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets

Vijeta Sharma [ID][a,b], Manjari Gupta [ID][a], Anil Kumar Pandey[c], Deepti Mishra [ID][d], and Ajai Kumar[b]

[a]Computer Science, DST- Center for Interdisciplinary Mathematical Sciences, Institute of Science, Banaras Hindu University, Varanasi, India; [b]Centre for Development of Advanced Computing, Pune, India; [c]Computer Centre, Banaras Hindu University, Varanasi, India; [d]Software, Data, and Digital Ecosystems Research Group, Educational Technology Laboratory, Department of Computer Science (IDI), NTNU - Norwegian University of Science and Technology, Gjøvik, Norway

ABSTRACT

Different types of research have been done on video data using Artificial Intelligence (AI) deep learning techniques. Most of them are behavior analysis, scene understanding, scene labeling, human activity recognition (HAR), object localization, and event recognition. Among all these, HAR is one of the challenging tasks and thrust areas of video data processing research. HAR is applicable in different areas, such as video surveillance systems, human-computer interaction, human behavior characterization, and robotics. This paper aims to present a comparative review of vision-based human activity recognition with the main focus on deep learning techniques on various benchmark video datasets comprehensively. We propose a new taxonomy for categorizing the literature as CNN and RNN-based approaches. We further divide these approaches into four sub-categories and present various methodologies with their experimental datasets and efficiency. A short comparison is also made with the handcrafted feature-based approach and its fusion with deep learning to show the evolution of HAR methods. Finally, we discuss future research directions and some open challenges on human activity recognition. The objective of this survey is to give the current progress of vision-based deep learning HAR methods with the up-to-date study of literature.

Introduction

Video surveillance has become a vital need in the smart city era to enhance the quality of life and develop the area as a safe zone. Surveillance cameras are usually installed at a certain distance for the proper coverage of an area. Therefore, better analysis and more in-depth understanding of videos are highly required, profoundly impacting the security system. A video data-driven system also helps healthcare, transportation, factory, schools, malls,

CONTACT Deepti Mishra ✉ deepti.mishra@ntnu.no 🖃 Software, Data, and Digital Ecosystems Research Group, IDI Educational Technology Laboratory, Department of Computer Science (IDI), NTNU – Norwegian University of Science and Technology, Teknologivegen 22, Gjøvik 2815, Norway

marts, etc. The objective of every camera feed is to know the specific incidence, such as identifying suspicious activities (Chen et al. 2020) at the airport, bus stop, railway station, unusual activities at public gathering events (S. Wang et al. 2021), an unusual pattern followed by the workers in the factory (Tao et al. 2018). These are the few exemplary areas where human activity recognition is highly desirable. In HAR-based system, usually, an alert generates to the control room for unusual activities. It is essential to know certain defined things in such cases instead of sitting in front of the camera feed and watching what is happening in every second of the camera feed.

Human activity recognition's primary objective is to accurately describe human actions and their interactions from a previously unseen data sequence. It is often challenging to accurately recognize humans' activities from video data due to several problems like dynamic background and low-quality videos. In particular, two main questions arise among various human activity recognition techniques: "Which action is performed?" – which comes under the action recognition task, and "Where exactly in the video?" is the localization task. The sequences of images are referred to as frames. Thus, the primary objective of an action recognition task is to process the input video clips to recognize the subsequent human actions.

Human activity mimics their habits; therefore, every human activities are unique, which turns into a challenging task to recognize. Moreover, developing such a deep learning-based model to predict human action within adequate benchmark datasets for evaluation is another challenging task. With the ImageNet (Jia Deng et al. 2009) dataset's immense success for image processing, several benchmark action recognition datasets (Kay et al. 2017; Soomro, Zamir, and Shah 2012) have also been released to pursue research in this area. Similarly, suppose we compare video data processing with image processing; it requires enormous computation power and a large number of input parameters to train the deep learning model.

Types of HAR System

There are two main categorizations of the HAR system based on the equipment:

Vision-based HAR

Static cameras installed at various places for surveillance purpose record the videos and store at servers. These camera feeds or recorded videos are then used for monitoring purposes. For example, (Htike et al. 2014) performed human posture recognition for video surveillance applications using one static camera. This type of HAR is used for road safety, public security, traffic management, crowd monitoring, etc. Figure 1 shows the typical steps of a vision-based human activity recognition system.

Figure 1. A typical human activity recognition system. (Image sequence (*) from Southampton database http://www.gait.ecs.soton.ac.uk/. Accessed: 2016–10-12).

Sensor-based HAR

Smartphones have become a global communication tool and, more recently, a technology for studying humans. Built-in sensors of smartphones can capture continuous information about human activities. Wan et al. (S. Wan et al. 2020) performed the identification of human activity using smartphone sensors. In this approach, data is retrieved from the smartphone's in-built accelerometer and gyroscope sensors, and then machine learning techniques were applied to recognize human activity. This type of HAR is helpful for patient monitoring systems, an individual player's activity monitoring during sports, etc., but cannot be applied to the broad application of human activity recognition for security at home/public places, monitoring, etc.

Motivation

Many factors have motivated us to perform this study. However, few significant factors are listed below:

- In ImageNet 2012 Challenge, we observed that a novel architecture of convolutional neural network (CNN) called AlexNet (Krizhevsky, Sutskever, and Hinton 2012) emerged as pivotal research in image processing and has proven to be a catalyst for video processing using CNN. Now researchers are focusing on deep learning-based HAR from real-time video processing.
- Video processing researchers are paying attention to develop deeper networks while utilizing GPU's harness with large training parameters. Thus, many deep learning methods have been developed for HAR, which must propagate among upcoming researchers.
- An adequate number of HAR video datasets is also attractive to the computer-vision community to set benchmarking methodologies. Thus, we are motivated to introduce the readers to the newest deep learning techniques for HAR and their evaluation of benchmark HAR video datasets.

Contribution

The current survey aims to provide the literature review of vision-based human activity recognition based on up-to-date deep learning techniques on benchmark video datasets. These video datasets are containing the video clips recorded from the static cameras installed at specific fixed locations. This paper presents the review in two portions – first introducing a benchmark video dataset and then giving the state-of-the-art HAR techniques in detail. The presentation of deep learning methods is the primary focus. Simultaneously, we also discuss the handcrafted feature-based approach and combined approach briefly to show the evolution in HAR techniques. In the previous literature surveys, researchers proposed a variety of taxonomy to categories the HAR methods. We show the novelty of our paper by introducing a taxonomy of deep learning-based HAR methods based on their network architecture and categorize each method under these categories. Also, the unique highlights of our study, which make it different from other surveys; are as follows:

- We comprehensively included almost all the advanced deep learning models shown in the literature up-to-date, outperforming the human activity recognition task.
- We present the performance mapping of each deep learning model with their experimented benchmark dataset.
- We show the evolution of action recognition from the traditional approaches to the advanced approaches.
- This paper aims to help the researchers of this field get detailed HAR information with techniques and datasets.

Furthermore, the objective behind choosing the time range of literature review between January 2011 and May 2021 for this study is that many deep learning architectures have been invented in this duration. We include few articles containing the handcrafted feature-based HAR methods only for background study, which is out of this time range. We chose much literature published in this duration and considered only those papers that were published only in Scopus and WoS's peer-reviewed journals.

   The rest of the paper is arranged as follows: section 2 shows the related literature surveys in this area, and section 3 lists the benchmark video datasets. Section 4 gives the detail of HAR methods, which presents various deep learning techniques. Section 5 elaborates the analysis of our study. Section 6 listed the significant challenges faced by the researchers. Section 7 shows the future directions, and section 8 concludes with this study.

Related Surveys

Human activity recognition techniques start from the handcrafted feature-based approach to advanced AI-based deep learning techniques. In related surveys, authors (Vrigkas, Nikou, and Kakadiaris 2015) have surveyed human activity recognition by dividing its scope into data modalities and their applications; in further sub-categories, the study divides based on the model development methods and various HAR activities. In the main categorization, the authors examine the unimodal and multimodal methods of HAR. In Unimodal categories, space-time, stochastic, rule-based, and shape-based models are grouped. Simultaneously, multimodal lists the affective, behavioral, and social networking sub-categories of human activity.

Reining et al.(Reining et al. 2019) performed a systematic literature review of HAR for production and logistics. This survey presents a detailed overview of state-of-the-art HAR approaches along with statistical pattern recognition and deep architectures. This study is beneficial for industrial applications. Beddiar et al. (Beddiar et al. 2020) surveyed vision-based human action recognition and categorized the entire study into the following fields: Handcrafted-feature and feature learning-based approach, where authors discussed the various techniques, including their implementation details. The authors also highlight related literature based on human activity types – Elementary human actions, Gestures, Behaviors, interactions, group actions, and events, which advocate HAR approaches at the minute level. Similarly, Zhu et al. (Zhu et al. 2016) also examined both handcrafted and learning-based approaches for action recognition. Unlike (Beddiar et al. 2020), the authors first evaluated the limitation of the handcrafted method then shows the rise of deep learning techniques of HAR in brief, till 2016. A review by Zhang et al. (S. Zhang et al. 2017) focuses on the advancement of state-of-the-art activity recognition approaches in terms of activity representation and HAR classification techniques. This survey categorizes the representation elements according to global, local, and advance depth-based, whereas categorization of classification techniques is based on template, discriminative, and generative models. The briefly explained models with the HAR dataset show performance accuracy in experimental results. All the HAR classification methods include in this study are till 2017.

Another survey by Herath et al. (Herath, Harandi, and Porikli 2017) of the same year shows a similar study, initiated with the pioneer of the HAR technique – a handcrafted-feature-based approach to deep learning-based methods. This survey is the first to present deep learning methods comprehensively, mapping with HAR datasets, which is missing in the previous surveys. But it includes the literature till 2016; therefore, advances after this

need to be present to researchers. The future direction mentioned by these authors is well defined, which is an excellent motivation to implement within the research community.

The survey conducted by Koohzadi et al. (Koohzadi and Charkari 2017) investigate the role of deep learning in image and video processing for the HAR. The overall approach is categorized into five types of models – Supervised-deep generative, Supervised-deep discriminative, Unsupervised deep, Semi-supervised deep, and Hybrid. One unique point highlighted in this survey is the benefits and tips & tricks of choosing a deep learning model for HAR in the abovementioned five categories. The author also discussed deep learning approaches to Spatio-temporal representation in terms of adding time as the third dimension of traditional image processing of 2D. Nweke et al. (Nweke et al. 2018) present a comprehensive review of deep learning methods for mobile and wearable sensor-based HAR. Categorization of methods is generative, discriminative, and hybrid by explaining the advantage and disadvantages. This study evaluates the deep learning methods on mobile sensor-based human activity recognition datasets, not vision-based activity recognition datasets. Authors also make the comparison of deep learning feature representation methods with conventional feature learning. Challenges of the HAR using sensor networks are also discussed.

The survey by Zhang et al. (H.-B. Zhang et al. 2019) shows progress in action features for depth and RGB data, advances in human-object interaction recognition methods, and recent deep learning-based action feature representation methods. This survey's main work comprehensively explains the handcrafted action feature for RGB, depth, and skeleton data, making this survey different from previous work. Deep learning methods performance evaluations are also discussed well with HAR datasets, which were presents till 2018. A survey was conducted by Singh et al. (Singh and Vishwakarma 2019) to guide researchers to match the suitable HAR dataset for benchmarking their algorithms. The existing HAR dataset categorizes into RGB and RGB-D(depth). Challenges with these datasets are also discussed in terms of illumination variation, annotation, occlusions, view variation, and fusion modalities. Liu et al. (B. Liu et al. 2019) presented the RGB-Depth sensor-based HAR survey where handcrafted and learning-based features are discussed. This survey shows a novel taxonomy for both the methods under three subcategories – Depth-based methods, Skeleton-based methods, Hybrid feature-based methods. This survey examines the deep learning method's accuracy performance on RGB-D-based human action datasets very briefly. Challenges and future research are also discussed for RGB-Depth sensor-based HAR. In the survey by Hussain et al. (Hussain, Sheng, and Zhang 2020), the authors discuss different areas of HAR with the primary focus on device-free techniques, especially RFID. The author proposes a new taxonomy based on the

related literature into three sub-areas: action-based, motion-based, and inter-action-based. These areas are further divided into 10 sub-topics and presented the latest HAR methods under each sub-area.

In a related survey by Dang et al. (Minh Dang et al. 2020), the authors comprehensively presented both the sensor-based and vision-based HAR methods. Further, each group divides into subgroups that perform different procedures, including the data collection, pre-processing methods, feature engineering, and the training process. A short description is also given for deep learning HAR methods along with challenges and future direction. Wang et al. (Lei Wang, Huynh, and Koniusz 2020) utilize the kinetics-based literature, which talks about Kinect camera for data collection and deep learning algorithms for activity recognition. The authors surveyed 10 Kinect-based algorithms for cross-subject action recognition and cross-view action recognition using six kinetics-based datasets. This survey is a novel source for researchers to develop HAR models for real-time applications[1] using Microsoft Azure Kinect Developer Kit[2] .The authors Jegham et al. (Jegham et al. 2020) addressed the challenges of HAR methods and datasets. They focused on surveying to investigate an overview of the existing methods as per the various types of issues shown in the literature. This survey motivates computer-vision researchers to find out the critical challenges in HAR to choose future research to overcome these real-world applications. To improve the accuracy of the HAR system, a survey conducted by Majumder et al. (Majumder and Kehtarnavaz 2021) reported the literature evidence of fusion of vision and inertial sensing. The study investigated in the form of fusion approaches, features, classifiers, and multimodality datasets is chosen, which is the first type of survey in this field. In a recent survey by Özyer et al. (Özyer, Ak, and Alhajj 2021), the authors categories the existing HAR methods into network-based, motion-based, multiple instances learning-based, dictionary-based, and histogram-based approaches. Also, they compared the results on HAR datasets.

Author (Verma, Singh, and Dixit 2022) has conducted a review on various supervised and unsupervised machine learning techniques for human behavior recognition. In this review, the authors reported the impactful literatures for abnormal behavior and activity recognition in the categories of supervised learning methods (classification and regression) such as support vector machine (SVM), Hidden Markov Model (HMM), and neural network. Whereas, in the category of unsupervised learning method (Clustering) for anomaly detection, author reported object trajectory analysis and pixel-based features for abnormal behavior detection in video sequence. Various types of clustering algorithms such as Partition-based clustering, hierarchical, density-based latent, Gaussian method for the applications of track analysis, moving hands, Multiple objects, behavior analysis, walking, running, and cycling on the highway.

To sum up, we observed that most of the surveys introduce a taxonomy for HAR methods categorization for comparison purposes. In the compared surveys, we also noticed a wide variety of HAR approaches: dataset-based, input-type-based, HAR real-world challenge-based, or learning-based. In this context, we state that the approach we have taken in this survey is the learning-based approach and propose a novel taxonomy of study based on the architecture of recent deep learning algorithms. We have covered more than 25 up-to-date deep learning-based algorithms and presented their performance on benchmark vision-based HAR datasets as reported in their original work. Table 1 summarizes the comparison of the surveys mentioned above. It includes the focus of study, taxonomy, year of publishing, source of data collection in those papers, whether deep learning-based HAR methods are included in brief or comprehensively, and a view on whether deep learning-based HAR methods are mapping with HAR datasets.

Benchmark HAR Datasets

In simple words, benchmarking is a way of discovering what the best performance being achieved is. The action recognition's benchmark video datasets are carefully prepared, validated, annotated, and earned good accuracy compared with its contemporary datasets. We list many action recognition datasets that are appeared in the various top-level international conferences[3] as a baseline for the action recognition algorithm challenge. Those algorithms achieved higher accuracy on these datasets and were top-ranked. Therefore, these datasets are known as benchmark datasets. This study includes the following benchmark video datasets to investigate various deep learning techniques for human action recognition and accuracy.

Ucf 101

The Center for Computer Vision Research, University of Central Florida, the USA, prepared the UCF101 human activity recognition dataset (Soomro, Zamir, and Shah 2012) in 2012. UCF101 is the newest version of the previously created UCF50 dataset, which contains 50 action categories. It has 13,320 videos of different categories of human actions, originally trimmed from YouTube. The UCF101 dataset is a widely adopted benchmark for action recognition, where all the activities are divided into 25 groups for each action category. This dataset is having 101 action classes and a minimum of 100 video clips in each class.

Table 1. Summary of previous HAR reviews included in this study.

Paper	Year	Source of Data	Focus of Study	Taxonomy	HAR Deep Learning Methods (In Brief/ Comprehensive)	Accuracy Mapping with HAR Benchmark Dataset
(Michalis, Nikou, and Kakadiaris 2015)	2015	Vision-based	Use of data from different modalities	Uni-model and Multimodal	Brief	No
(Zhu et al. 2016)	2016	Vision-based	Overview of recent advancements in HAR	Handcrafted and Learning-based action representations	Brief	Yes
(S. Zhang et al. 2017)	2017	Vision-Based	Highlights the advances of state-of-the-art HAR approaches	Activity representation and classification methods	Brief	Yes
(Herath, Harandi, and Porikli 2017)	2017	Vision-Based	Overview of the important steps taken toward HAR	Handcrafted and Deep Learning	Comprehensive	Yes
(Koohzadi and Moghadam Charkari 2017)	2017	Image + Video data	Deep learning-based HAR	Extending Spatial CNN, separately temporal structure, Modeling temporal sequences	Brief	Yes
(Nweke et al. 2018)	2018	Sensor-based	In-depth summaries of deep learning methods for mobile and wearable sensor-based HAR	Generative, Discriminative and Hybrid methods	Brief	No
(Reining et al. 2019)	2019	Sensor-Based	Marker-based Motion Capturing in Logistics and Production	Shallow classification methods	Brief	Yes
(H.-B. Zhang et al. 2019)	2019	Vision-based	Depth analysis of feature representation and action recognition with different aspects	Human action feature representation, action recognition, human-object interaction recognition, human action detection	Comprehensive	Yes
(B. Liu et al. 2019)	2019	RGB- Depth camera	Human action and interaction analysis	Handcrafted and Deep Learning	Brief	No
(Zawar, Sheng, and Emma Zhang 2020)	2020	Device-Free - Radio Frequency Identification (RFID) technology	Explore various areas of HAR on device-free solutions, with main focus on RFID technology	Action-based, Motion-based, and interaction based.	Brief	No
(Minh Dang et al. 2020)	2020	Vision and Sensor-based	Analysis of vision and sensor-based HAR approaches	Data collection, Feature extraction, HAR common methodology	Brief	Yes

(Continued)

Table 1. (Continued).

Paper	Year	Source of Data	Focus of Study	Taxonomy	HAR Deep Learning Methods (In Brief/ Comprehensive)	Accuracy Mapping with HAR Benchmark Dataset
(Lei, Huynh, and Koniusz 2020)	2020	Vision-based (Kinetics camera)	Analysis and comparison of recent Kinect-based algorithms	cross-subject action recognition and cross-view action recognition	Comprehensive but kinetics based deep learning HAR methods	Yes, On Kinetics based datasets
(Beddiar et al. 2020)	2020	Vision-based	Review HAR Approaches, datasets and classify activity recognition process	Feature extraction process, recognition system, input data modalities, supervision level	Brief	No
(Jegham et al. 2020)	2020	Vision-based	Real-world challenges and solution to vision-based sensors, an overview of action recognition	Based on challenges-	Brief	No
(Majumder and Kehtarnavaz 2021)	2021	Depth + inertial sensing + RGB	Comparing vision and inertial sensor fusion approaches and multimodal datasets	Single-Modality and multi-Modality	Brief	Yes
(Özyer, Duygu Selin, and Alhajj 2021)	2021	Vision-Based	Overview of different HAR approaches based on principles and datasets	Network, Motions, Multiple Instance Learning, Dictionary, Histogram	Brief	Yes
(Verma, Singh, and Dixit 2022)	2022	Vision-Based	Supervised and Unsupervised Machine Learning Approach	Various classification, regression and clustering methods	Comprehensive	No
Our Approach	2021	Vision-Based	Comprehensive recent deep learning-based HAR methods on benchmark datasets	Based on the design and network architecture of deep learning methods for HAR	Comprehensive	Yes

Hmdb 51

Brown University released the HMDB51 (H Kuehne et al. 2011) dataset, where most of the videos are from movies, and some are from public databases and online video libraries such as YouTube. The database contains 6849 samples, divided into 51 categories, and each class has at least 101 samples. The frame rate of clips is 30 fps.

JHMDB

Joint-annotated Human Motion Data Base JHMDB (Jhuang et al. 2013) is a fully annotated data set for human actions and human poses released by the Brown University research group. It contains 923 videos, which are categorized into 21 different activities, namely pick, run, stand, sit, brush hair, pour, throw, shoot the bow, catch, clap, wave, shoot, climb stairs, pull-up, golf, push, jump, kickball, shoot ball, gun, swing baseball, walk. This dataset is a comparatively less used dataset among all.

Kinetics

The Kinetics 400 dataset (Kay et al. 2017) was released by Deepmind, focusing on human actions (rather than activities or events). With more than 400 clips for each class, each from a unique YouTube video, are two orders of magnitude larger than previous datasets, HMDB-51 and UCF-101. Categories of several action classes are: i) Person Actions (singular) – drawing, drinking, punching, laughing, etc. ii) Person–Person Actions – kissing, hugging, shaking hands, etc., and iii) Person-Object Actions- mowing the lawn, opening gifts, mopping, washing dishes, etc. The dataset contains 400 human action classes, and each class has 400 or more clips, each from a unique video, for a total of 240k training videos. The length of each clip is around 10 s. The test set is also having 100 clips for each class. An extension of the Kinetics human action dataset from 400 classes to 600 classes is available as Kinetics 600 (Carreira et al. 2018), released in 2018. Further, a dataset with 700 action classes was released in 2019 as kinetics 700 (Carreira et al. 2019), all by Deepmind. The vision for the Kinetics dataset is that it becomes the ImageNet equivalent of video data.

Hollywood-2

Hollywood-2 (Bojanowski et al. 2014) is a human actions and scenes dataset with 12 classes of human actions and 10 classes of scenes distributed over 3669 video clips and a total of 787720 frames containing sequences from 69

Hollywood movies. A set of video clips, each one is annotated as a sequential list of actions, such as "walking" then "sitting," then "answer the phone." The length of a total video is 20.1 hours.

Breakfast Dataset

The breakfast dataset (Hilde Kuehne, Arslan, and Serre 2014) consists of 10 classes related to breakfast preparation performed by 52 different individuals in 18 different kitchens. The dataset is considered one of the biggest fully annotated human action datasets. Video clips of this dataset are recorded "in the wild" instead of recording the data in a controlled lab environment.

It appears closer to real-world conditions as it behaves as it is monitoring and doing analysis of daily activities. The total length of the video is approximately 77 hours. The cameras used to record human actions are webcams, standard industry cameras, and a stereo camera. All videos were downsampled to a resolution of 320 × 240 pixels with a frame rate of 15 fps. Here cooking activities include preparing sandwiches, tea, coffee, orange juice, chocolate milk, a bowl of cereals, pancakes, fried eggs, fruit salad, and scrambled egg. This dataset was released by SERRE Lab, Brown University, in 2014.

Charades

Charades (Sigurdsson et al. 2016) comprises of daily indoor activities. Videos are recorded as acting out the sentence like in a game of Charades. It is one of the most extensive public datasets with continuous action videos, containing 9848 videos of 157 classes (7985 training and 1863 testing videos). Each video is ~30 seconds. Because of its different variety of activities and long-duration clips, it is a challenging dataset. This dataset is released by The Allen Institute for Artificial Intelligence in 2016.

AVA

AVA (Gu et al. 2018) dataset consists of 80 different unique visual actions, with 57.6k video segments collected from approximately 192 movies. Video clips are 3 s-long and extracted sequentially in 15-min chunks from each movie. Using a batch of 15 min per video enables variations simultaneously in the video. A total of 210k actions are labeled. Google released it in 2018, and it appeared at the conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Epic-Kitchens-55

Epic Kitchen-55 (Damen et al. 2018) is the largest dataset in first-person (egocentric) vision, released by the University of Bristol, UK, and the University of Catania, Italy, in April 2018. It has 39596 action segments recorded by 32 individuals while performing routine daily activities in their kitchen environments. Each action is labeled as a combination of a verb and a noun in this dataset, e.g., "cut vegetable," "wash utensil," etc. There are 331 noun classes and 125 verb classes, but still, these are heavily imbalanced. Total video is 55 hours long in full HD format. The latest Epic kitchen version is available with 100 hours of full HD video with 45 individual kitchen recordings, Epic-Kitchen-100, and released in July 2020.

Something – Something

The 20BN-SOMETHING-SOMETHING dataset (Goyal et al. 2017) is an extensive collection of densely labeled video clips that show humans performing pre-defined basic actions with everyday objects. The total Number of Videos is 220847, and the total number of classes is 174. The JPG images were extracted from the original videos at 12 frames per second. TwentyBN released it in June 2017. The specialty of the dataset is that a large number of crowd workers created it.

Moments-in-Time

The Moments in Time (MIT) dataset (Monfort et al. 2020) is a large-scale video classification dataset. Its objective is to help AI systems recognize and understand actions and events in videos. It has more than 800 K videos (~3 seconds per video). It is released by CSAIL, MIT, in 2018. The moment-in-time recognition challenge appeared in CVPR, 2018, and later on ICCV 2019 as Moments in Time Multimodal Multi-Label Action Detection Challenge.

ActivityNet

ActivityNet (Heilbron et al. 2015) is a large-scale video benchmark dataset for human activity understanding. It covers a wide range of complex human activities that interest people in their daily living. ActivityNet provides video clips of 203 activity classes. ActivityNet has 849 hours long video and was released in 2016 as version −2. In contrast, the older version was released in 2015 and contained 100 activity classes. This dataset illustrates three scenarios to compare algorithms for human activity understanding: global video classification, trimmed activity classification, and activity detection.
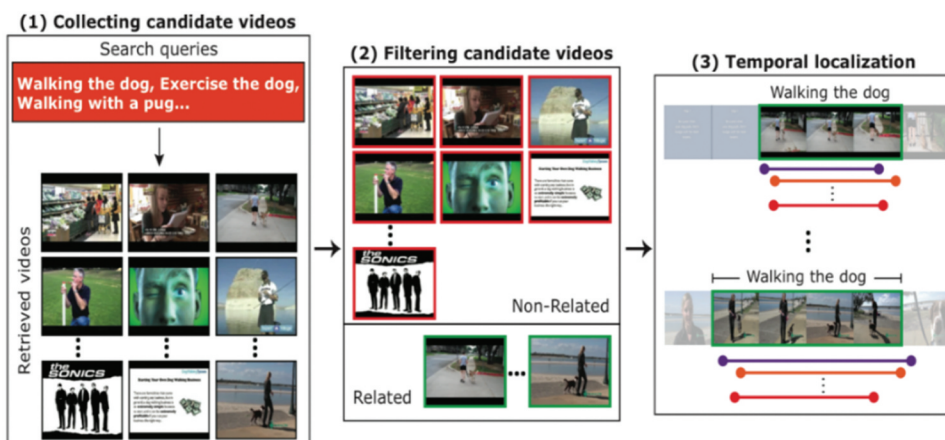
Figure 2. Three steps of ActivityNet human activity collection and annotation process (Source: (Heilbron and Carlos Niebles 2014)).

Data were annotated with the help of Amazon Mechanical Turk (AMT). Figure 2 shows the three main steps of data acquisition: Collection, Filtering, and Temporal Localization. In step 1, candidate videos were searched on the web for each ActivityNet category by multiple queries. Step 2 shows the filtration process, where collected videos from step 1 were verified by Amazon Mechanical Turk workers and deleted videos unrelated to any exciting activity. In step 3, every video was provided by temporal localization, where starting and ending times were marked. The action is performed, and annotation was done to every video clip by AMT. The annotation files are stored in JSON format. This dataset first appeared in CVPR, 2015.

Sports- 1 M

The Sports-1 M dataset (Karpathy, Toderici, Shetty, Leung, Sukthankar, and Li 2014b) consists of roughly 1.2 million YouTube sports videos annotated with 487 classes, and it is representative of videos in the wild. Approximately 1000 to 3000 videos are available for each class. And, around 5% of the videos are annotated with more than one class. However, this dataset is the largest publicly available sports action video dataset, but the annotations that it gives are at the video level. No detailed information is mentioned about the location of the class of interest. Google released this dataset in 2014. This dataset first appeared in CVPR, 2014. Table 2 shows the list of the above benchmark datasets.

Table 2. List of benchmarks HAR datasets.

Dataset	No. of Classes	No. of Video Clips	Description	Source of Data Collection	URL	Release Year	Paper
UCF 101	101	13320	Realistic action videos, an extension of the UCF50 data set which has 50 action categories	YouTube	https://www.crcv.ucf.edu/data/UCF101.php	2012	(Soomro, Roshan Zamir, and Shah 2012)
HMDB 51	51	6849	Each class containing a minimum of 101 clips	Movies, public databases and YouTube	https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/#dataset	2011	(H Kuehne et al. 2011)
JHMDB	21	928	2D pose annotation for scale, pose, segmentation, coarse viewpoint, and dense optical flow	21 action categories extracted from HMDB51	http://jhmdb.is.tue.mpg.de/dataset	2013	(Jhuang et al. 2013)
Kinetics400	400	300000	YouTube video URLs dataset	YouTube video	https://deepmind.com/research/open-source/kinetics	2017	(Kay et al. 2017)
Kinetics600	600	500000	YouTube video URLs dataset	YouTube video	https://deepmind.com/research/open-source/kinetics	2018	(Carreira et al. 2018)
Kinetics700	700	65000	YouTube video URLs dataset	YouTube video	https://deepmind.com/research/open-source/kinetics	2019	(Carreira et al. 2019)
Breakfast Dataset	10	1989	consists of 10 cooking activities performed by 52 different actors in multiple kitchen locations	Manually Recorded	https://serre-lab.clps.brown.edu/resource/breakfast-actions-dataset/	2014	(Hilde Kuehne, Arslan, and Serre 2014)
Charades	157	9848	Indoor Activities	Amazon Mechanical Turk (AMT)	https://prior.allenai.org/projects/charades	2016	(Sigurdsson et al. 2016)
AVA	80	57600	spatio-temporal localization of atomic visual actions	192 different movies	https://research.google.com/ava/	2018	(Chunhui et al. 2018)
Epic-Kitchens	149	432	action segmentation dataset	Manually recorded	https://epic-kitchens.github.io/2018	2018	(Damen et al. 2018)
Something-Something	174	220847	basic actions with everyday objects	Recording by crowd workers	https://20bn.com/datasets/something-something	2017	(Goyal et al. 2017)
Moments in Time - Dataset	339	1000000	capturing visual and/or audible actions	Web-Search	http://moments.csail.mit.edu/	2020	(Monfort et al. 2020)
Sport 1-1 M	487	1200000	sports action video dataset	YouTube	https://deepai.org/dataset/sports-1m#__sid=js0	2014	(Karpathy et al. 2014b)

(Continued)

Table 2. (Continued).

| Dataset	No. of Classes	No. of Video Clips	Description	Source of Data Collection	URL	Release Year	Paper
ActivityNet	200	20000	Videos for Human Activity Understanding	Web-Search	http://activity-net.org/download.html	2015	(Heilbron et al. 2015)
Hollywood Extended	16	937	Segmentation and classification of actions computed as mean over frames	Movies	https://www.di.ens.fr/willow/research/actionordering/	2014	(Bojanowski et al. 2014)
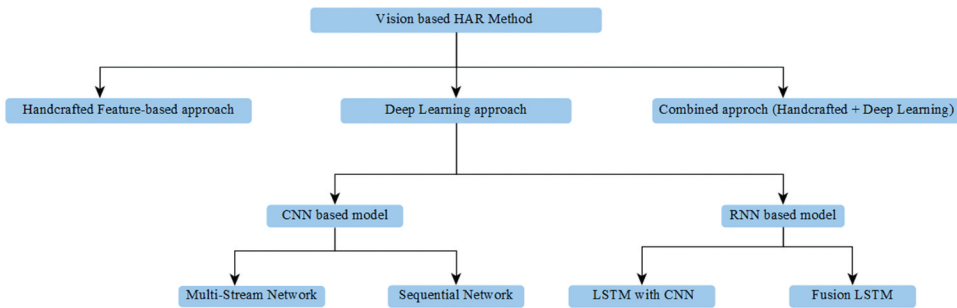
Figure 3. Proposed taxonomy of vision-based HAR Methods.

Vision-based HAR Methods

Human activity recognition is a challenging time series classification task. In a vision-based approach, it involves predicting the movement of a person by analyzing the camera feeds. The problem of action recognition in videos can vary widely, and there's no single approach that suits all the problem statements. In this section, we present the various techniques of HAR, right from the traditional approaches to the advanced Deep learning approaches. Thus, we have divided this section into three sub-sections – i) Handcrafted-feature-based approach, ii) Deep learning-based approach, and iii) Fusion approach. This paper's primary focus is to present a deep learning-based approach mainly; We demonstrate this approach in-depth, whereas the other two sub-sections are brief. Therefore, we further divide the deep learning-based approach into CNN and RNN-based approaches. Categorization of methods based on network architecture shows the novelty of our survey. Figure 3 shows the proposed taxonomy of vision-based HAR methods, on which basis we organize the entire study.

Handcrafted feature-based Approach

This section investigates how human activity recognition was done before deep learning, using a handcrafted feature-based approach. This approach usually includes a three-stage process – feature extraction, feature classification, and feature representation. Handcrafted feature extractors such as HOG (Dalal et al. 2005), HOF (Laptev et al. 2008), SIFT (Lowe 2004), SURF (Bay, Tuytelaars, and Van Gool 2006), etc., are used to extract the low-level features. Further, to make the final feature classification, some specific feature representation algorithms encode them into global feature-based methods and local feature-based methods. The global feature treats the representation as a whole. The region of interest is generally located by tracking or background subtraction; after that, this region is considered as a whole representation. The local

representation shows the observation as a group of independent features offered collectively or at space-time points where the location belongs to exciting motions. Since more specific feature representations enhance classification performance, rewriting the original feature with discriminative words is vital to this HAR method. Authors have proposed various feature representation algorithms, and most of them have shown effective performance in feature encoding.

The older methods (Kläser, Marszałek, and Schmid 2008; Scovanner, Ali, and Shah 2007) implemented for human activity recognition in video data mainly focus on handcrafted features. These methods primarily considered motion and appearance information while using various local features. Local features have already been proven a successful technique for the image recognition task. Local features describe images in terms of Speeded Up Robust Features (SURF) (Bay, Tuytelaars, and Van Gool 2006) and Scale-Invariant Feature Transform (SIFT) (Lowe 2004). After the massive success of image recognition techniques, various researches have been done by directly extending the image classification methods to learn Spatio-temporal information of video for action recognition. Kläser et al. (Kläser, Marszałek, and Schmid 2008) developed Histograms of Oriented 3D spatiotemporal Gradients (HOG3D) by extending the HOG descriptor for action recognition in video.

Similarly, an extension of SIFT, a new technique SIFT-3D (Scovanner, Ali, and Shah 2007), was proposed to extract the spatiotemporal motion features for action recognition. One practical approach is Dense Trajectories (DT) (H. Wang et al. 2013a), which consist of HOG, HOF, and Motion Boundary Histogram (MBH). The same author (H. Wang and Schmid 2013) further added some new features based on temporal templates with dense trajectories, which are called combinedly as "Improved Dense Trajectories." They constructed templates by considering a video sequence as a third-order tensor and computing three different projections. Further, they used several functions to protect the fibers from the video sequences and combined them through sum pooling.
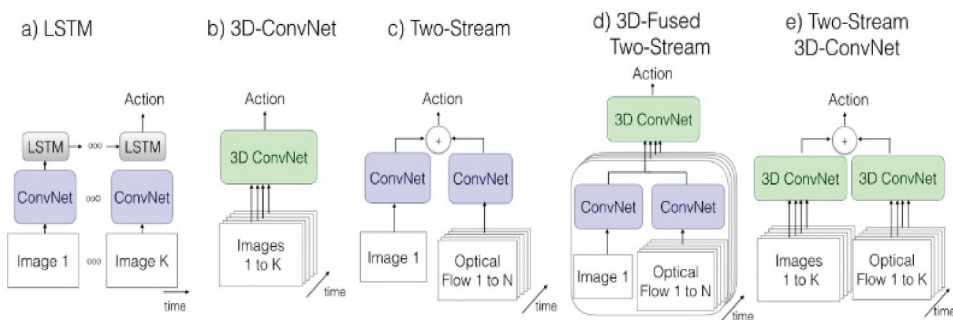

Figure 4. Deep learning-based HAR architectures. K shows the total number of frames in the input video, and N shows the subset of neighboring frames of the video (Carreira and Zisserman 2017).

Unfortunately, these feature-based approaches have some drawbacks. First of all, most of the commonly used feature extractors are developed based on a specific dataset, and the feature extractors are often database-biased. They do not have general-purpose feature extraction ability. And secondly, creating a handcrafted feature-based human activity recognition system required careful feature engineering. Thus, handcrafted feature engineering works are labor-intensive and time-consuming; severely hindering the development of related technologies. Therefore, newer HAR systems are deep learning-based techniques.

State-of-the-art Deep Learning Approach

In the past few years, deep learning techniques (Krizhevsky, Sutskever, and Hinton 2012; H. Wang et al. 2013b; Simonyan and Zisserman 2015a; Szegedy et al. 2015) outperformed on image data with the ImageNet dataset's advent. Many deep learning algorithms have been developed for human activity recognition from video data, especially on the benchmark datasets (Kay et al. 2017; Soomro, Roshan Zamir, and Shah 2012; H Kuehne et al. 2011; Sigurdsson et al. 2016; He et al., n.d.). Unlike handcrafted-feature-based approaches, a deep learning-based HAR method can simultaneously learn visual features, feature representations, and classifiers. Deep learning architectures have different variants, but the most attractive model for vision-based HAR is Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), which have achieved very promising results on benchmark video datasets. Videos can be considered as a sequence of individual images. Therefore, many deep learning practitioners quickly treat video classification as performing image classification techniques as a sum of all the frames. In this section, we present state-of-the-art deep learning techniques for human activity recognition. In the deep learning approach, we have further sub-categorized this method as a CNN-based and RNN-based system. Figure 3 shows the categorization of CNN-
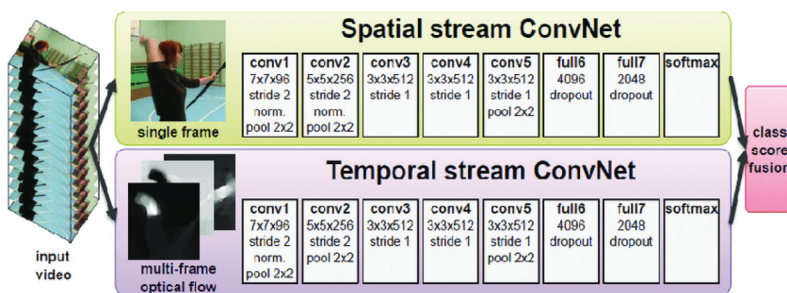

Figure 5. An example of multi-stream CNN architecture (Source Simonyan et al. (Simonyan and Zisserman 2014)).

based methods as – multi-stream networks and sequential networks, whereas RNN-based methods as – LSTM with CNN and Fusion LSTM. Figure 4 shows the different video processing architectures for HAR methods. For the input as RGB video, the convolutional layers use 2D image-based and optical flow layers, whereas 3D (video-based) layers include pre-computed optical flow. Another is combinations of 2D ConvNets and temporally recurrent layers LSTMs. In Figure 4,5 architectures (c), (d), and (e) are categorized as CNN-based multi-stream networks, whereas (b) type architectures are placed inside CNN-based sequential networks. Also, architecture (a) shows the RNN-based LSTM with CNN architecture, whereas RNN-based models that use more than LSTM with CNN are grouped into the FusionLSTM approach.

Convolutional Neural Network (CNN) Based Approach

In the past few years, CNNs have been proven to be one of the most successful image processing models. The deeper architecture of CNN is helpful for object recognition from static images (Krizhevsky, Sutskever, and Hinton 2012). But the traditional CNN architectures are not suitable for video processing. Karpathy et al. (Karpathy et al. 2014b) experimented with video action recognition using stacked video frames as input to the network. The results were unsatisfactory and even lower than handcrafted feature-based approaches (H. Peng et al. 2016; Wang and Schmid 2013). Further, the researchers also propose various architectures by using different benchmark action datasets. This paper further categorized these advanced CNN-based methods into a multi-stream network and a sequential network, based upon the model's architecture.

Multi-Stream Network

This section categorizes deep learning models that use separate spatial and temporal streams in CNN architecture for HAR. Figure 4(c,d,) shows the concept of multi-stream CNN architectures. A novel, widely used approach, introduced by (Peng et al. 2016), two-stream CNN, where first stream is for video's spatial feature, whereas the second stream focuses on temporal features. The spatial stream recognizes the action from still images, and the temporal stream performs action recognition in the form of dense optical flow. Finally, these two streams are combined using late fusion – this methodology of action recognition is proven to be better than handcrafted-based methods (H. Wang and Schmid 2013). Similar work was extended to action recognition from videos in (Karpathy et al. 2014b), using stacked video frames as input to the network, but the results were worse than the previous approach. Later on, researchers relied on the fact that two-stream architecture is not applicable for human activity recognition from live camera feeds in the real-time scenario because of its computational complexity. In (Feichtenhofer,

Pinz, and Zisserman 2016) (the extension of (Simonyan and Zisserman 2014)), authors fuse the spatial and flow streams after the last network convolutional layer, showing some improvement on the HMDB51 dataset while requiring less test time augmentation (snapshot sampling). Followed by this method, Carreira et al.(Carreira and Zisserman 2017) implemented a technique by using Inception-V1. The spatial and temporal streams passed through the 3D convolutional layer before going to the last average pooling layer of Inception-V1.

In another work, inspired by the two-stream ConvNets (Simonyan and Zisserman 2014), the authors propose a novel three-stream ConvNet model (C. Di Huang, Wang, and Wang 2016). This network introduces an additional movement ConvNet stream along with spatial ConvNet and Temporal ConvNet. The purpose of the movement ConvNet stream is to distinguish the action which has similar pose change but with different speed or direction; for example, walking versus running and push versus pull. The movement ConvNet handles these kinds of actions. The input of the moving stream is the centroid of detected human regions. Finally, these three streams are combined with the hinge loss classifier to classify human activity correctly. Similar work was also performed in (Najeera, Anu, and Sadiq 2018), where "Movement Stream" is the third stream; in contrast, spatial and temporal streams perform the same functionality as in (Feichtenhofer, Pinz, and Zisserman 2016). This additional movement stream is a traditional neural network (Simonyan and Zisserman 2015b), which takes the same vector of the human centroid (Huang et al. 2016) and combines these three scores streams using a hinge loss classifier in the last layer. A long-range temporal-based network was introduced by (Limin Wang et al. 2016). This network aims to utilize the visual information of entire videos to perform video-level prediction. This model is built on top of the (Simonyan and Zisserman 2014). In this method, learning of CNN on video data with the help of temporal segment networks and limited training data was proven to be a very effective method compared to its contemporary methods of HAR. This network overcomes frame-level prediction issues (Karpathy et al. 2014b) by predicting the action for long-range videos. The authors of (Feichtenhofer, Pinz, and Wildes 2017) introduce an architecture based on multiplicative interactions of space-time features called SpatioTemporal Network. This model is the combination of motion and appearance in the form of a two-stream. The multiplicative gating functions have also been evaluated for the residual networks and closely observed the effect of the gating function on model accuracy. This model also injects an identity mapping function to track the long-term dependencies and learn temporal features. This architecture is fully convolutional in space-time and can evaluate a video for human action recognition in a single forward pass.

Hao et al. (Hao and Zhang 2019) proposed a model for HAR from video data called Spatiotemporal Distilled Dense-Connectivity Network (STDDCN). This network is partially inspired by (G. Huang et al. 2017), which uses a dense-connectivity and knowledge distillation network. This model focuses on exploring the relationship between appearance and motion streams along with multiple features. The dense network explicitly enhances the relationship of the Spatio-temporal features at the feature representation layers. In addition to that, knowledge distillation among these two streams and last fusion allows both streams to communicate with the last layers. The unique architecture of STDDCN enhances the capacity of the model to acquire high-level ordered spatiotemporal features.

In (Verma and Singh 2021) author proposed a fusion method using two 3D Convolutional Neural Network (3DCNN) and a Long Short Term Memory (LSTM) network from RGB, Depth and skeleton joint positions, then score generation using trained SVM model, thereafter, fusion of score and optimization using Evolutionary algorithms – Genetic algorithm (GA) and Particle Swarm Optimization (PSO) algorithm. Thus, accuracy achieved were with GA 85.93% and with PSO 83.75%.

Author (Verma, Brij Mohan Singh, and Chauhan 2020) proposed a 2D CNN-based algorithm to recognize single-limb and multi-limb human activities. In the first step, single-limb and multi-limb activities are separated, then the separated single- and multi-limb activities recognize using sequence-classification. Thus, they achieved the overall accuracy as 97.88%.

Sequential Network

Figure 4(b) depicts the network architecture of CNN-based sequential networks, or we can say those networks, uses a single stream or stacked structure. This type of 3D ConvNet architecture is like a natural approach to video modeling and is just like standard convolutional networks, but with Spatio-temporal filters. This type of network structure has specific characteristics – they directly create hierarchical representations of Spatio-temporal data.

To overcome the problem of computing motion information of frames, Ji et al. (Ji et al. 2013) proposed a 3D convolution in which features are computed from spatial and temporal at the convolution stages of CNNs. Unlike the conventional 2D CNNs, where convolutions are applied to compute features from the spatial dimensions on the 2D feature maps, the 3D convolution network is modeled by convolving a 3D kernel by keeping multiple sequential frames together in the cube format. Thus, the convolution layer's feature maps connect to the network's back layer contiguous sequential frames. Thereby capturing motion information in 3D CNN, kernel extracts one feature from the frame, while kernel weights apply in the entire cube. Instead of using 2D convolutions across frames, Tran et al. (Tran et al. 2015) use a 3D convolution network on video data in their work. This network train the 3D CNN on the

Sports-1 M dataset and then use them (or an ensemble of nets with different temporal depths) as a feature extractor for other video datasets. Afterward, they utilize a simple linear classifier, such as a support vector machine on top of the extracted features. Thus, it has proven to be more accurate than the other state-of-the-art algorithms. Further, it has been observed that this model could have performed even better if handcrafted features like HOG, HOF, or IDT were used additionally. In this experiment, five random two-second clips were extracted from every video as an action performed in the entire video at the training time. And during testing, 10 clips were randomly sampled, and predictions made across them were averaged for final prediction.

Varol et al. (Varol, Laptev, and Schmid 2018) present a new architecture of a two-stream convolutional neural network with Long-short-term Spatiotemporal Features (LSF CNN). This network aims to recognize human action from video data fast and efficiently compared to previous networks. This complete network is a fusion of two subnetworks. The first subnetwork is a long-term spatiotemporal features extraction network (LT-Net), which receives the RGB frames as inputs. Another subnetwork is the short-term spatiotemporal features extraction network (STNet) that accepts the optical flow data as input. Further, these two streams fuse in the CNN fully connected layer. Finally, the fully connected layer's output sends to the simple classifier support vector machine (SVM). This model includes a novel approach for better utilizing the optical flow field, which has better performance than CNN-based deep learning models (Feichtenhofer, Pinz, and Zisserman 2016). They followed conventional methods to use optical in action recognition problems. This model can learn very deep features in both spatial and temporal areas in this fusion-based two-stream network.

Zheng et al. (Zheng and Zhang 2020) introduce a cross-modal architecture that includes an "Alignment Network" and a "Fusion Network" to improve the performance of the human activity recognition as compared to a model that uses three separates streams of CNN (Huang et al. 2016; Simonyan and Zisserman 2014). In this model, the first step is to extract the different modal information mapped into a common subspace to align. After that, the aligned features are further combined to generate correlated, consistent, and complementary representations. In the last layer, the learned features are fed as input for actual action recognition to the classifier. The MARS method (Crasto et al. 2019) proposes two learning approaches to train a standard 3D CNN. It operates on a single RGB frame that explicitly mimics the motion stream. Thus, it saves the optical flow computation cost at test time. Two learning approaches perform in this model are – i) by minimizing a feature-based loss compared to the Flow stream, the network reproduces the motion stream with high fidelity, ii) leverages the effect of appearance and motion information simultaneously. The model is trained with a linear combination of

the feature-based loss and the standard cross-entropy loss for action recognition in this method. As a single stream, MARS performs better than RGB or Flow alone.

Yang et al. (Yang et al. 2018) propose an Asymmetric 3D CNN model that works on asymmetric single-direction 3D convolution architecture to assess the conventional 3D convolution network (Shuiwang et al. 2013). In this model, the Asymmetric 3D convolutions network enhances the capacity of feature learning. This model is a collection of local 3D convolutional networks, called MicroNets, which are built by incorporating multi-scale 3D convolution branches. To efficiently perform the action recognition task, an asymmetric 3D-CNN deep network is developed using these MicroNets. Unlike those models (Simonyan and Zisserman 2014), trains separately for two–streams network. This model minimizes the training efforts for the RGB frames and optical Flow frames separately. A simple but effective multi-source enhanced input is also implemented here, where vital information of the RGB and optical flow frames are fused at the early preprocessing stage. Moreover, the performance of this model is evaluated on a benchmark dataset. This asymmetric 3D-CNN model outperforms all the conventional 3D-CNN models. Its accuracy for action recognition is also compared with various state-of-the-art CNN-based human action recognition models on benchmark datasets.

Principal Component Analysis Network (PCANet), proposed by Abdelbaky et al. (Abdelbaky and Aly 2020), selects a subset of frames from each action. At the same time, a feature vector is calculated from the previously trained PCANet for each frame. All feature vectors are then fused, and their dimensionality is reduced by using the Whitening Principal Component Analysis algorithm (WPCA)(Thameri et al. 2011). The Support Vector Machines (SVM) classifier is used in the output layer, followed by the block-wise histograms for the feature pooling layer. Thameri et al. (Thameri et al. 2011) compare this HAR method to observe the impact of using features from the first convolutional stage, and this network is named as PCANet-1, whereas using deep features from the second stage is named as PCANet-2.

RNN Based Approach

There is a problem with the approach, where the video considers a sequence of images and processes them to combine for classification. Video classification is more than just a simple image classification. There is an assumption that videos have subsequent frames (images) in which they correlate to semantic content. In this way, if videos' temporal nature can include, it improves actual video classification performance. Neural network architectures, such as Recurrent Neural Networks (RNNs) are suited for time series data. Particularly, Long Short-Term Memory (LSTMs) have outperformed on

video data for human action recognition. But practically, these methods are very resource-hungry and time-consuming for training over thousands of video files.

However, inspired by the success of RNN in sequential information modeling (Sutskever, Vinyals, and Le 2014), many researchers propose an LSTM model for action recognition to learn spatiotemporal features with CNN effectively. We explain these techniques under the 'LSTM with CNN' approach. Research shows that advanced RNN architectures have also been developed, which uses more than LSTM with CNN. We categorize these methods under the 'fusion LSTM' approach. The description of HAR methods under these two RNN-based approaches is given in the following sub-sections:

LSTM with CNN

Srivastava et al. (Srivastava, Mansimov, and Salakhutdinov 2015) use multi-layer Long Short-Term Memory (LSTM) networks to learn high-level features of video sequences. An encoder LSTM employ in this model to map the input sequence for a fixed-length representation. This representation is decoded by using a single or multiple LSTMs decoder to output the numerous tasks, such as predicting the future sequence, reconstructing the input sequence, etc. The author shows two types of input sequences in this model – First, patches of image pixels and, secondly, "percepts." These are high-level representations of video frames and extracted through a pre-trained CNN model.

A Lattice-LSTM architecture, proposed by Sun et al. (L. Sun et al. 2017), is an extension of LSTM by learning hidden state transitions of memory cells for individual spatial locations. This method effectively and efficiently increases model dynamics' ability across time and focuses on the non-stationary long-term motion dynamics issue without significantly increasing the model complexity. This network works differently than traditional two-stream architectures(Crasto et al. 2019; Najeera, Anu, and Sadiq 2018). Lattice-LSTM uses RGB and optical flow to train input gates and forget gates in one network. Traditional two-stream architecture considers these two data as separate entities without information of each other. This architecture reduces the complexity of the method by avoiding multiple streams.

Li et al. (Li et al. 2018) propose a novel architecture for end-to-end learning of human actions in video data, called VideoLSTM. Instead of training the input video with the unique features of recurrent or convolutional neural network architectures, VideoLSTM architecture builds according to the input video's essential requirements. Initiating the VideoLSTM functionality from the soft-Attention LSTM, this model took advantage of the spatial correlation and used a convolution layer in the soft-Attention LSTM architecture. Also, the authors introduce motion-based attention. Another architecture proposes in (C. Dai, Liu, and Lai 2020) utilizes the visual attention mechanism and introduces an end-to-end two-stream attention-based LSTM

network. This network aims to overcome the problems of assigning the same weights on different visual and temporal cues in the parameter training stage by most CNN-based HAR network architecture(Feichtenhofer, Pinz, and Zisserman 2016; C. Dai, Liu, and Lai 2020; Baccouche et al. 2011). This problem mainly affects the feature distinction determination. This model can selectively focus on the compelling features for the original input images and pay different attention levels to each deep feature map's outputs. A correlation layer is proposed to adjust the deep learning network parameter based on the correlation judgment by considering the relation between two feature streams. Here, two-stream attention-based LSTM consists of two streams, such as temporal feature stream and spatial-temporal feature stream. For the temporal stream, the authors design this network to automatically determine the dominant area in every optical flow-based image by temporal attention module; also, it sums the representation values of these images to make a feature vector. On the other hand, an LSTM model is used for the spatial-temporal stream after the pooling layer to effectively learn the spatial maps' unique temporal relationship. In this entire process, spatial-temporal attention is assigned different weights for the different levels of features. In the end, a joint optimization layer is used to optimize the loss for the deep model to achieve reasonable accuracy.

Fusion LSTM

Few RNN architectures use very deep CNN and LSTM networks. We place those types of architectures of the HAR method in this section. Baccouche et al. (Baccouche et al. 2011) introduce an idea of fully automated deep learning architecture, which trains to classify human actions without using any prior knowledge. The basis of this model is – converting a vanilla CNN model to a 3D network, which automatically learns Spatio-temporal features. Further, an RNN-LSTM architecture trains for the classification by feeding the temporal feature of the already learned pattern w.r.t. every timestep. This model's comparative result has proven to be more efficient than the existing CNN-based approach for human activity recognition.

Before this technique, various researches have been done based on using LSTMs on separately trained feature maps to observe if temporal information can be captured from clips. Unfortunately, researchers conclude that convoluted features temporal pooling had been proven more effective than LSTM stacked after trained feature maps. In (Donahue et al. 2017), the authors design a network on the same idea of using LSTM blocks (decoder) after convolution blocks (encoder) but using end-to-end training of the entire architecture. They compared RGB and optical flow as input choices and observed that a weighted scoring of predictions based on both inputs was the best. Here During training, 16 frame clips are sampled from the video. The RNN-based architecture is trained with input as RGB or optical flow of 16 frame clips. The final

prediction for each clip is based on the average of predictions of each time step. Thus, the final prediction at the whole video level is the average of predictions calculated from each clip. Even though the end-to-end training frameworks were available in this model. This model still has a few drawbacks, such as false label assignment, inability to capture long-range temporal information, and more efforts for pre-computing optical flow.

Ng et al. (Ng et al. 2015) evaluate the numerous deep learning architectures to merge individual image information in the entire video for more extended periods than the (Donahue et al. 2017). This paper proposes two methods in this architecture, which can process the long-length videos very efficiently. In the first method, the authors apply a multiple convolutional temporal feature pooling network. It is useful, especially when implementing CNN for the Two-Stream LSTM network architecture. The second method is built on the generic concept of considering a hierarchical sequence of frames as video data. This model uses two CNN architectures to process individual video frames: AlexNet and GoogLeNet, which stacks inception modules to form a network 22 layers deeper that is substantially different from the previous CNNs model. A deep LSTM architecture has been used in this model, in which the output from one LSTM layer is input for the next layer. The authors experimented with various layers and memory cells and used five stacked LSTM layers, at last, each with 512 memory cells. This architecture is one of the most popular method for HAR.

Wang et al. (X. Wang et al. 2019) propose a primarily decomposed model into two modules: Three Dimension Inception (I3D) network and Long Short-Term Memory (LSTM) work. In this model, I3D architecture extracts spatial features and captures motion features in sequential frames. Further, the output feature trained by the I3D model serves as the input of the LSTM network, which is mainly responsible for modeling high-level spatial features. As a result, video features were learned with high efficiency and represented in low-level and high-level features. To avail the benefit of pre-training, authors pre-trained the 3D CNN model on the Kinetics dataset to improve the model's generality. And then, long short-term memory (LSTM) is introduced to model the high-level temporal features produced by the Kinetics pre-trained 3D CNN model.

Wan et al. (Y. Wan et al. 2020) present a two-stream CNN (Simonyan and Zisserman 2014) with LSF- CNN-based new architecture. This complete network is a fusion of two subnetworks. The first subnetwork is a long-term spatiotemporal features extraction network (LT-Net), which receives the RGB frames as inputs. Another subnetwork is the short-term spatiotemporal features extraction network (STNet) that accepts the optical flow data as input, which calculates from the two subsequent frames. Further, these two streams are fused in the CNN fully connected layer, and finally, the fully connected layer's output is sent to the simple classifier support vector machine (SVM). This model includes a novel approach for better utilizing the optical flow field,

which has better performance than CNN-based deep learning models (Feichtenhofer, Pinz, and Zisserman 2016). They followed conventional methods to use optical flow in action recognition problems. This model can learn very deep features in both spatial and temporal areas in this fusion LSTM architecture.

Zhao et al. (Zhao and Jin 2020) develop a novel convolutional and recurrent network for action recognition, which is "doubly deep" in spatial and temporal layers. Therefore, in the feature extraction stage, an improved p-non-local operation as a simple and effective component was introduced to capture long-distance dependencies with deep convolutional neural networks. Similarly, the class prediction stage combines a Fusion KeyLess Attention with the forward and backward bidirectional LSTM to learn the data's sequential nature more efficiently and elegantly, using multi-epoch models fusion based on confusion matrix. The most important parts are the p-non-local block developed to solve the long-distance dependencies to reduce the computational complexity and the Fusion KeyLess Attention with Bi-LSTM, which aims to pay attention to the part of the interest of human motion. The CNN+p-non-local module takes the video frames as inputs and produces feature maps X. The hidden state of bidirectional LSTM obtains the KeyLess Attention Mask. Next, the model computes the feature map weighted according to the attention mask at each timestep and finally sends it to LSTMs for prediction. Yu et al. (Yu et al. 2020) introduce a new long-term temporal feature learning network for human action classification called "Pseudo Recurrent Residual Neural Networks (P-RRNNs)." This network uses the recurrent model and performs the fusion of each in the different connections among all the units. P-RRNNs primarily use two-stream CNNs architecture – GoogLeNet, to extract local temporal and spatial features, respectively. These local deep features are further combined into global long-term temporal features with the help of two-stream P-RRNNs.

Combined Approach

Some researches show the deep learning architectures combined with feature-based techniques for HAR. This progress is shown in (Hao et al. 2018), where the author presents an Asymmetric 3D-CNN architecture in their extended research, and they fed the Asymmetric 3D CNN model with the enhanced RGBF frames. Thus, the deep features from RGB and RGBF frames are further combined with the traditional improved Dense Trajectories (IDT) (H. Wang and Schmid 2013) features used to train a linear SVM to classify actions. The resulting Asymmetric 3D-CNN (RGB+RGBF+IDT) out-performs the newest state-of-the-art HAR methods on the benchmark dataset. Duta et al. (Duta et al. 2017) show the Improved Dense Trajectories (IDT) approach to utilize the handcrafted features by keeping the default parameter settings recommended to extract four different descriptors: HOG, HOF, MBHx., and MBHy.

Table 3. List of HAR methodologies.

Approach	Architecture [Model]	Description	Experiment-al dataset	Accuracy (in %)	Paper	Publish Year
Hand-Crafted feature-based approach	Improved Dense Trajectories + Support Vector Machine		UCF 101, HMDB 51	83.8, 52.1	Action recognition by dense trajectories (H. Wang et al. 2011)	2011
	Improved dense trajectories + Optical Flow + RGB		UCF101, HMDB 51, KTH, Weizmann	89.3, 65.3, 97.5, 98.8	Human action recognition by means of subtensor projections and dense trajectories (Carmona and Climent 2018)	2018
CNN based approach	2 Stream ConvNet	CNN-M-2048	UCF 101, HMDB 51	88.1, 59.4	Two-Stream Convolutional Networks for Action Recognition in Videos (Simonyan and Zisserman 2014)	2014
	Multiresolution CNN	Context stream + fovea stream	UCF 101	63.3	Large-scale Video Classification with Convolutional Neural Networks (Karpathy et al. 2014b)	2014
	Convolutional Two Stream	VGG-M-2048 + VGG 16	UCF 101, HMDB 51	92.5%, 65.4%	Convolutional Two-Stream Network Fusion for Video Action Recognition (Feichtenhofer, Pinz, and Zisserman 2016)	2016
	3 Stream CNN	(Pre-Trained VGG-16 CNN model) Spatial Stream + Temporal Stream + movement Stream	KTH	54.06 [Spatial] 59.50 [Temporal] 36.57 [Movement]	An Intelligent Action Predictor from Video using Deep Learning (Najeera, Anu, and Sadiq 2018)	2018
	3 Stream ConvNet	Spatial, Temporal, Movement	UCF 101, HMDB 51 [selected only 9 categories from both datasets]	93.42	Human Action Recognition System for Elderly and Children Care Using Three Stream ConvNet (Huang et al. 2016)	2016
	Temporal Segment Network	Spatial ConvNet + Temporal ConvNet	UCF 101, HMDB 51	94.2, 69.4	Temporal segment networks: Toward good practices for deep action recognition (Limin Wang et al. 2016)	2016
	SpatioTemporal Network	Two stream ConvNet (ResNet50 + ResNet 152)	UCF 101, HMDB 51	94.2, 68.9	Spatiotemporal multiplier networks for video action recognition (Z. Liu and Hu 2019)	2017

(Continued)

Table 3. (Continued).

Approach	Architecture [Model]	Description	Experiment-al dataset	Accuracy (in %)	Paper	Publish Year
	Spatiotemporal distilled dense-connectivity network (STDDCN)	DenseNet based knowledge distillation and dense-connectivity	UCF 101, HMDB 51	93.78, 67.52	Spatiotemporal distilled dense-Connectivity network for video action recognition (Hao and Zhang 2019)	2019
	Two-Stream Inflated 3D	based on 2D ConvNet inflation with pretrain on − 1. Only ImageNet[a] 2. ImageNet and Kinetics[b]	UCF 101, HMDB 51	[98.0,80.7][a], [97.8,80.9][b]	Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset (Carreira and Zisserman 2017)	2017
	3D CNN	Three streams of CNN	KTH	90.2	3D Convolutional Neural Networks for Human Action Recognition (Shuiwang et al. 2013)	2013
	C3D	ConvNet + linear SVM	UCF 101	82.3	Learning Spatiotemporal Features with 3D Convolutional Networks (Tran et al. 2015)	2015
	Cross-Model CNN	ResNet50+ sparse contractive auto-encoder (SCAE)+ Deep Belief Network (DBN)	UCF 101, HMDB 51	94.8, 68.5	A Cross-Modal Learning Approach for Recognizing Human Actions (Zheng and Zhang 2020)	2020
	Motion-Augmented RGB Stream	3D CNN + Motion Augmented RGB Stream	UCF 101, HMDB, Something Somethingv1	98.1, 80.9, 53.0	MARS: Motion augmented RGB stream for action recognition (Crasto et al. 2019)	2019
	Asymmetric 3D Convolutional Neural Networks	CNN + RGB + RGBF	UCF 101, HMDB 51	89.5, 63.5	Asymmetric 3D Convolutional Neural Networks for action recognition (Hao et al. 2018)	2018
	PCANet –	CNN+ Principle Component Analysis +SVM	KTH	80.49[PCANet-I], 87.52 [PCANet – II]	Human Action Recognition based on Simple DeepConvolution Network PCANet (Abdelbaky and Aly 2020)	2020
	Long-Term Temporal Convolutions (LTC)-CNN	LTC +Optical Flow + RGB	UCF 101, HMDB 51	91.7, 64.8%	Long-term Temporal Convolutions for Action Recognition (Varol, Laptev, and Schmid 2018)	2018

Table 3. (Continued).

Approach	Architecture [Model]	Description	Experiment-al dataset	Accuracy (in %)	Paper	Publish Year
RNN Based Approach	Two Stream Attention-based LSTM	LSTM	UCF11, iHMDB	89,6, 66.3	Human action recognition using two-stream attention-based LSTM networks(C. Dai, Liu, and Lai 2020)	2020
	Lattice LSTM	CNN + Lattice LSTM	UCF 101, HMDB 51	93.6, 66.2	Lattice Long Short-Term Memory for Human Action Recognition (Lin et al. 2017)	2017
	VideoLSTM	ConvNet + Attention LSTM	UCF 101, HMDB 51	89.2, 56.4	VideoLSTM Convolves, Attends and Flows for Action Recognition (Zhenyang et al. 2018)	2018
	Composite LSTM	Multilayer LSTM network	UCF 101	84.3	Unsupervised learning of video representations using LSTMs (Srivastava, Mansimov, and Salakhutdinov 2015)	2015
RNN Based Approach	Long-term Recurrent Convolutional Networks	CNN+LSTM	UCF 101	82.9	Long-term Recurrent Convolutional Networks for Visual Recognition and Description Visual Recognition and Description (Donahue et al. 2017)	2017
	Two-Stream LSTM	CNN (AlexNet + GoogleNet) + LSTM	UCF 101	88.3	Beyond Short Snippets: Deep Networks for Video Classification (Ng et al. 2015)	2015
	I3D-LSTM	Inflated 3D + LSTM	UCF 101	95.1	I3D-LSTM: A New Model for Human Action Recognition (X. Wang et al. 2019)	2019
	3D ConvNet LSTM	3D CNN + LSTM	KTH 1, KTH 2	94.39, 92.17	Sequential Deep Learning for Human Action Recognition (Baccouche et al. 2011)	2011
	Long-Term SpatioTemporal Features(LSF)-CNN	two-stream convolutional network with long-short-term spatiotemporal features	UCF101,HDMB51,	94.8, 70.2	Action Recognition Based on Two-Stream Convolutional Networks With Long-Short-Term Spatiotemporal Features (Wan et al. 2020b)	2020
	Attention CNN-RNN	VGG19 + 3 layer Bi-directional LSTM	HMDB 51, Hollywood2	50.1, 59.6	Human Action Recognition Based on Improved Fusion Attention CNN and RNN (Zhao and Jin 2020)	2020
	Pseudo Recurrent Residual Neural Networks(P-RRNNs)	GoogLeNet+ LSTM (IP-LSTM)	UCF 101, HMDB 51	88.5, 58.6	Learning Long-Term Temporal Features With Deep Neural Networks for Human Action recognition (Sheng et al. 2020)	2020
Combined Approach	Asymmetric 3D CNN (RGB + RGBF+IDT)	CNN + IDT	UCF 101, HMDB 51	95.6, 65.4	Asymmetric 3D Convolutional Neural Networks for action recognition (Hao et al. 2018)	2018

(Continued)

Table 3. (Continued).

Approach	Architecture [Model]	Description	Experiment-al dataset	Accuracy (in %)	Paper	Publish Year
	ST-VLAD (CNN + IDT)	VGG19 ConvNet + IDT	UCF50, UCF 101, HMDB 51	97.9, 91.5, 67.6	Spatio-Temporal VLAD Encoding for Human Action Recognition in Videos (Duta et al. 2017)	2017
	videoLSTM + IDT	ConvNet + ALSTM+IDT	UCF 101, HMDB 51	91.5, 63.0	VideoLSTM convolves, attends and Flows for action recognition (Zhenyang et al. 2018)	2018
	Temporal Vector of Locally Aggregated Descriptors (T-VLAD)	C3D + feature Descriptors + SVM	UCF101	89.0	T-VLAD: Temporal Vector of Locally Aggregated Descriptor for Multiview Human Action Recognition (Binte Naeem et al. 2021)	2021
	Hierarchical Feature Reduction & Deep Learning (HFR-DL)	CNN (AlexNet) + LSTM + HOG +KNN	DCF101	93.90	Complex Human Action Recognition Using a Hierarchical Feature Reduction and Deep Learning-Based Method (Serpush and Rezaei 2021)	2021

This approach is one of the state-of-the-art handcrafted approaches for feature extraction. Each of these four descriptors was extracted along all valid trajectories and combined with the Vector of Locally Aggregated Descriptors (VLAD) (Jégou et al. 2012), a popular super vector method. Further, the authors experimented with the Spatio-temporal VLAD (ST-VLAD) model, an extended encoding method that includes Spatio-temporal data at the encoding stage. This is achieved by using a video sequence and pulling specific information from every video clip.

Later, inspired by (Duta et al. 2017), Naeem et al. (Binte Naeem et al. 2021) propose a new method T-VLAD for an efficient multiview human action recognition. It uses simple C3D convolutional features that learn the long-term temporal pattern of the input video. Further, it comprises the spatial location of local features and the temporal location of global features frame by frame. This network architecture successfully extends the frame-level information into complete video-level information. T-VLAD uniquely recognizes the human action by distinguishing the human motion. This method works sturdy for action recognition in the variable background scene.

Serpush et al. (Serpush and Rezaei 2021) considered the complexity of the preprocessing phase of previously developed models (Duta et al. 2017; Sheng et al. 2020) and proposed a model architecture that automatically chooses similar frames from the input video. They retrieve the primary features of only selected frames instead of extracting whole features. This hierarchical architecture combines background subtraction, HOG, deep neural network, and skeletal modeling method in the preprocessing step. The entire network is consisting of CNN, and the LSTM, which works as a feature selector, and finally, a Softmax-KNN is employed for the classification task.

In the VideoLSTM model (Zhenyang et al. 2018), a complementary approach was followed by using IDT features for human activity recognition. For IDT, the authors used the software from (H. Wang and Schmid 2013) and the implementation of the classification pipeline. The author observes VideoLSTM result combines with IDT features improves the HAR performance.

Table 3 shows the List of HAR Methods along with its benchmark dataset(s) and accuracy.

A Quantitative Analysis

In this section, we provide a high-level analysis of the HAR datasets and methods as mentioned above. We highlight the performance of some remarkable HAR methods by focusing on the following point-method performance, architectural complexity, commonly used approaches, and the popularity of datasets. We present the analysis from two ways:

HAR Dataset Analysis

HAR datasets are a vivid variety of qualities based upon their parameters, such as RGB, RGB-D(Depth), Multiview, recorded in a controlled environment. Other parameters are – recorded "In the wild," annotated with a complete sentence, annotated with only action label datasets, etc. We list all these varieties of datasets in Table 2 with their comprehensive information, such as the source of data collection, number of actions, video clips, nature of datasets, and released year to show the progress in this area. We observe that most of the HAR datasets could not become a popular choice among computer-vision researchers due to their over simplicity, small size (Jhuang et al. 2013; Bojanowski et al. 2014; Hilde Kuehne, Arslan, and Serre 2014), and unsatisfactory performance(L. Sun et al. 2017). However, there is no such thing as the most accurate standard datasets, i.e., on which researchers measure the HAR method to set as a benchmark, but of course, as we observe UCF101(Soomro, Roshan Zamir, and Shah 2012) and HMDB51 (H Kuehne et al. 2011) are the dominating datasets for researchers interest. The reason reported by Serpush et al. (Serpush and Rezaei 2021) for choosing UCF101 over other datasets is that it contains well categorization of actions, which helps to train the models efficiently. Also, the actions played in the recorded clips are, by various individuals, while in other datasets, the activities and actions are usually performed by one actor only.

We found enough gap in the works of literature despite the emerging of newer datasets in the last 5 years (Carreira et al. 2018, 2019; Damen et al. 2018; Goyal et al. 2017; Heilbron and Niebles 2014; Kay et al. 2017; Sigurdsson et al. 2016). Researchers have chosen UCF101 (Soomro, Roshan Zamir, and Shah 2012) and HMDB51 (H Kuehne et al. 2011) datasets very frequently. We notice YouTube URL-based datasets (Carreira et al. 2018, 2019; Kay et al. 2017) seem a bit complex to fetch the clips while training. This training process requires additional computational power and processing time. Despite a well-annotated dataset (Monfort et al. 2020), it still seems heavily imbalanced, which might be a reason to attract the HAR researchers less. Also, researchers are bound for a limited choice of availability of Sports datasets (Karpathy et al. 2014b; Lin et al. 2017). In Dataset Hollywood Extended (Sigurdsson et al. 2016), clips are extracted from movies recorded from several viewpoints; this brings multiple-viewpoint variations to the video streams. We note that Hollywood 2 extended and SPORTS-1 M datasets are well annotated, but labeling is slightly noisy.

We observe that ActivityNet (Heilbron et al. 2015) dataset is prepared to reduce previous datasets' limitations and enhance the possibility of training and tuning large networks.

HAR Methods Analysis

Our study includes a list of HAR and results on benchmark datasets reported in the literature. It is shown in Table 3. We analyze these methods with the main focus on their network architecture. On this basis, we introduce novel taxonomy. Broadly, we categories them into Handcrafted-feature-based, Deep learning-based, and combined approach. We further examine the deep learning – CNN & RNN-based HAR method's architectures exquisitely followed by combined approach in brief.

Focusing on handcrafted methods, dense trajectory descriptors (H. Wang et al. 2013a) show satisfactory results on UCF101. Moreover, these descriptors can easily incorporate with optical flow for temporal feature pooling(Carmona and Climent 2018). At this point, from the above-mentioned handcrafted architecture, we analyze that dense trajectory descriptors (Carmona and Climent 2018) which outperform the deep learning CNN-based solutions (Simonyan and Zisserman 2014) on UCF101 with 89.3% and 88.1, respectively.

Convolutional Two Stream (Feichtenhofer, Pinz, and Zisserman 2016) network shows significant performance on UCF 101 (92.5%) and HMDB 51 (65.4%) from its previously designed architectures (Feichtenhofer, Pinz, and Zisserman 2016) due to having two separate streams – spatial and temporal in the form of CNN. A step ahead, by adding one more stream (movement stream), the architecture is converted into three streams (Najeera, Anu, and Sadiq 2018), which gave an average result on the KTH dataset; however, the authors did not report performance on other datasets.

Further improvement, by introducing 3D architecture for HAR method (Tran et al. 2015), still shows the same accuracy as two-stream until the existence of Inflated 3D architecture – reported 98.0%,80.7% accuracy on UCF101 and HMDB51, respectively. Surprisingly, this architecture got the highest accuracy on the HMDB51 dataset on previously introduced methods. Auto-encoder-based CNN architecture (Zheng and Zhang 2020), motion augmented RGB stream architecture (Crasto et al. 2019), and Asymmetric 3D CNN architecture (Hao et al. 2018) have almost similar accuracy despite having diversity in the complex internal architecture. Principle Component analysis-based CNN architecture (Abdelbaky and Aly 2020) implicitly utilizes the Whitening Principal Component Analysis method and the block-wise histograms for the feature pooling layer. This compound network makes this CNN architecture completely different than others, and the evaluated result is 87.5% on the KTH dataset.

A network designed by considering video data as time-series, LSTM based HAR methods have various architecture reported in the literature included in our work. Turning our attention to RNN-based LSTM architectures, we find that the LSTM network (Donahue et al. 2017; Ilya, Vinyals, and Le 2014; Taylor et al. 2010; Thameri et al. 2011) learns spatiotemporal features more effectively than

CNN, which also outperforms. The work in (X. Wang et al. 2019) also suggests deeper architecture than (Carreira and Zisserman 2017) like models, which helps to improve the performances. To lead in competitive results on the HMDB-51 and UCF-101 datasets, LSTM architecture (Lin et al. 2017; Zhenyang et al. 2018) achieves equal accuracy (84.5%), whereas (Ng et al. 2015; Srivastava, Mansimov, and Salakhutdinov 2015) reported more than 90% accuracy. Lately, both the CNN and RNN structures include 3D convolution filters (Wan et al. 2020b). Attention-based LSTM architecture uses a three-layer Bi-directional LSTM approach (Zhao and Jin 2020) but a comparatively lower result on HMDB51 and Hollywod2 datasets than (Ilya, Vinyals, and Le 2014; Thameri et al. 2011). In Combined architecture(Crasto et al. 2019; Lin et al. 2017; Sheng et al. 2020), handcrafted and deep learning HAR methods formed a deeper network and gave spirited performance (above 90% accuracy) on UCF101, HMDB51 datasets. This indicates that the HAR method architecture developed by deep networks is complementary to the handcrafted approach. In most cases, we must mention that both deep learning networks and trajectory descriptors consider similar inputs – RGB and optical flow frames and observed that combining them gave the best result among all approaches.

Challenges

Although there has been significant progress over the past few years, there are still many challenges in applying Deep Learning models to build vision-based action recognition systems and bring their benefits to our lives. Various challenges for the human activity recognition task are low-quality videos because of long-distance, multisubject interactions and group activities and complex and changing backgrounds, intraclass variation, and interclass similarity. In addition, the lack of a dataset is still a challenge. More HAR video datasets from various domains are highly required. Designing the algorithm for real-time recognition of human activity from multi-camera systems and camera-calibration are still a big challenge.

Future Research

Human activity recognition, a thrust area of video data processing research and new approaches, regularly solves HAR issues with the above-discussed advanced deep learning techniques. However, a study conducted by Jegham et al. (Jegham et al. 2020) shows that few significant challenges are still a huddle in video processing, such as background clutter, high computational complexity, responsiveness to illumination changes, fast irregular motion, occlusion, viewpoint changes, etc. Several advanced research are required to overcome these challenges. Many HAR future research topics and breakthrough techniques that are interesting to investigate are listed below:

Deep Learning Hybrid Network

A hybrid network structure of existing deep learning-based methods can be developed as a holistic approach to boost the feature extraction technique. It is considered that each algorithm is developed for specific feature extraction and learning purposes. Therefore, the tuning of few prominent methods can be useful in diverse HAR applications, rather than developing a technique for some particular dataset, restricted domain, and application. This hybrid network may also be helpful to tackle the significant issues of HAR, such as – camera rotation angles, cluttered background, and occlusions.

Human Behavior Analysis

Human behavior analysis is a less explored area of HAR research. Detecting abnormal behavior and predicting subsequent activity based upon behavior could be the possible research area. This would be very useful for public security, surveillance, and other application for detecting abnormal action. For example, Sun et al. (B. Sun et al. 2021) prepared a dataset of students' behavior in the classroom environment and designed an algorithm for behavior analysis. The extension of similar work in other domains will give a novel direction to the researchers. Another conceivable direction is to analyze logical and semantic relations between behaviors and activity, including natural language processing (NLP) techniques.

Recognizing Complex Activities

Most of the existing HAR methods recognize simple actions, such as eating, running, sitting, jumping, etc. But in a real scenario, these simple actions are part of some complex activities: eating pasta, cleaning glass, cooking vegetables, etc. Complex activities are a series of multiple actions performed in a systematic order. Future research can be designing systems for recognizing complex activities that are composed of simple actions. The key technology for complex action recognition research could be the combination of data modalities, such as RGB data, depth data, and skeleton data (Coppola et al. 2020).

Development of State-of-the-art HAR Dataset

The future direction of HAR research is motivated to develop a more realistic dataset. To overcome the challenges of real-time HAR, a new dataset could consider as a merger of existing homogenous datasets. Therefore, a complex model can consume a scale dataset to train, test, validate. However, every

dataset is considered a benchmark in its specific domain; uniting them would be more practical. Moreover, in the future, a standard parameter could be defined to annotate HAR datasets to fit with any model.

Efficient Transfer Learning

After the immense success of transfer learning in image processing, now emerging research is transfer learning for video processing (Chakraborty et al. 2021; Serpush and Rezaei 2021) on cross datasets, hyperparameter, fine-tuning, and utilization of pre-trained HAR models. A few pre-trained action recognition models like I3D pre-trained on ImageNet, and Kinetics400 datasets, SlowFast model, pre-trained on HMDB51/UCF101 datasets are freely available on MXNet[4] to ease the researchers. This practice is a time-saver approach as well as helpful in the non-availability of high computation power. To set up a benchmark in transfer learning for video–processing in on urge.

Data Augmentation

Lack of datasets is often a huddle in HAR research; therefore, data augmentation techniques give a way to move on with on-the-fly generated data. Generative Adversarial Network(GAN) (X. Dai, Yuan, and Wei 2021) got immense popularity among the computer-vision community. It provides sophisticated domain-specific data augmentation and generative solutions, such as video-to-video translation. Other data augmentations techniques[5] can be explored and match up with the HAR methods to train the model on small datasets.

Leveraging High-Performance Computing Power

Analyzing human action data requires significantly increased computational power than still images when deep learning techniques are employed. Achieving fast and accurate action detection is a critical issue in real-time applications. The convergence of HPC-AI technologies (Huerta et al. 2020) indicates an appropriate research direction to harness the power of super-computing to train a very deep network in a multi-GPU parallel computing platform for timely, quick, and accurate analysis in real-time. The high accuracy of real-time video processing is in high demand. These could be an important direction of future research on human action recognition – fast action detection in the spatiotemporal dimension. People perform actions differently and at different speeds; therefore, Real-time inferencing of algorithms is a must.

Conclusion

This review shows that much successful research has been done in human activity recognition on benchmark video datasets. Initially, applying CNNs at the frame level helped improve the HAR model's accuracy compared to the traditional handcrafted manual feature-based extraction techniques. Later on, 3D-CNNs improved the accuracy of CNN's by applying and processing a set of frames at a time. Many advanced HAR models started using RNNs and LSTMs to include the temporal component of the videos efficiently. The Two Stream Fusion method improved the performance over C3D without the extra parameters used in C3D. The TSN architecture has attempted to tackle two significant challenges in action recognition – overfitting due to small sizes and long-range modeling. However, extra computation power usage for pre-computing optical flow and related input modalities is still a problem. The choice of VLAD as an effective way of feature pooling was already proved long back. This model's extension at an end-to-end trainable framework proven this technique extremely robust and state-of-the-art for most action recognition tasks in early 2017. By using hidden Two-stream, there was an improvement in speed and associated cost of prediction. The authors showed the dependency on slower traditional methods to compute optical flow with the automated generation of optical flow. It was observed that using T3D architecture; the results didn't improve on I3D results, which can mainly be attributed to a much lower model footprint than I3D. The contribution of the model was the supervised transfer learning technique. All the papers included for review are from peer-reviewed journals, indexed in Scopus and Web of Science, and published between 2011 and 2021. However, the supporting evidence of related survey used to show the evolution of HAR methods (such as literature included for handcrafted feature-based approach) is older than 2011, and existing datasets for video processing are referenced arXiv too, such as(Carreira et al. 2018; Soomro, Roshan Zamir, and Shah 2012). We notice that most of the HAR methodologies developed even after the advent of big datasets like Kinetics, EPIC Kitchen, Something–Something, etc., have experimented on UCF 101 and HMDB51 datasets. In addition to this, we observe that HAR architectures achieved higher accuracy on UCF 101 than the HMDB51 dataset.

Endnotes

1. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp = &arnumber = 8197492
2. https://www.microsoft.com/en-us/p/azure-kinect-dk/8pp5vxmd9nhq?activetab = pivot %3aoverviewtab
3. International Conference on Machine Learning – https://icml.cc/, Computer Vision and Pattern Recognition – http://cvpr2021.thecvf.com/ etc.
4. https://cv.gluon.ai/model_zoo/action_recognition.html
5. https://github.com/okankop/vidaug

Disclosure statement

Funding

ORCID

Vijeta Sharma ⓘ http://orcid.org/0000-0002-5889-3652
Manjari Gupta ⓘ http://orcid.org/0000-0003-1939-5383
Deepti Mishra ⓘ http://orcid.org/0000-0001-5144-3811

References

Abdelbaky, A., and S. Aly. 2020. "Human Action Recognition Based on Simple Deep Convolution Network PCANet." *Proceedings of 2020 International Conference on Innovative Trends in Communication and Computer Engineering, ITCE 2020, Aswan, Egypt*, 257–62. doi:10.1109/ITCE48509.2020.9047769.

Baccouche, M., F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. 2011. Sequential Deep Learning for Human Action Recognition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7065 LNCS:29–39. doi:10.1007/978-3-642-25446-8_4.

Bay, H., T. Tuytelaars, and L. Van Gool. 2006. SURF: Speeded up Robust Features. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 3951 LNCS:404–17. doi:10.1007/11744023_32.

Beddiar, D. R., B. Nini, M. Sabokrou, and A. Hadid. 2020. Vision-Based Human Activity Recognition: A Survey. *Multimedia Tools and Applications* 79 (41–42):30509–55. doi:10.1007/s11042-020-09004-3.

Bojanowski, P., R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. 2014. Weakly Supervised Action Labeling in Videos under Ordering Constraints. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8693 LNCS:628–43. doi:10.1007/978-3-319-10602-1_41.

Carmona, J. M., and J. Climent. 2018. Human Action Recognition by Means of Subtensor Projections and Dense Trajectories. *Pattern Recognition* 81:443–55. doi:10.1016/j.patcog.2018.04.015.

Carreira, J., and A. Zisserman. 2017. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset." *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, Hawaii* 2017Janua: 4724–33. doi:10.1109/CVPR.2017.502.

Carreira, J., E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman. 2018. A Short Note about Kinetics-600. *ArXiv*. http://activity-net.org/challenges/2018/evaluation.html .

Carreira, J., E. Noland, C. Hillier, and A. Zisserman. 2019. A Short Note on the Kinetics-700 Human Action Dataset. *ArXiv: 1907.06987*.

Chakraborty, S., R. Mondal, P. Kumar Singh, R. Sarkar, and D. Bhattacharjee. 2021. Transfer Learning with Fine Tuning for Human Action Recognition from Still Images. *Multimedia Tools and Applications* 80 (13):20547–78. doi:10.1007/s11042-021-10753-y.

Chen, D., P. Wang, L. Yue, Y. Zhang, and T. Jia. 2020. Anomaly Detection in Surveillance Video Based on Bidirectional Prediction. *Image and Vision Computing* 98:103915. doi:10.1016/j.imavis.2020.103915.

Christoph, F., A. Pinz, and R. P. Wildes. 2017. "Spatiotemporal Multiplier Networks for Video Action Recognition." *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, Hawaii* 2017-Janua: 7445–54. doi:10.1109/CVPR.2017.787.

Christopher, R., F. Niemann, F. Moya Rueda, G. A. Fink, and M. ten Hompel. 2019. Human Activity Recognition for Production and Logistics-a Systematic Literature Review. *Information (Switzerland)* 10 (8):1–28. doi:10.3390/info10080245.

Chunhui, G., C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, L. Yeqing, and S. Vijayanarasimhan, et al. 2018. "AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions." In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*, 6047–56. doi:10.1109/CVPR.2018.00633.

Claudio, C., S. Cosar, D. R. Faria, and N. Bellotto. 2020. Social Activity Recognition on Continuous RGB-D Video Sequences. *International Journal of Social Robotics* 12 (1):201–15. doi:10.1007/s12369-019-00541-y.

Crasto, N., P. Weinzaepfel, K. Alahari, and C. Schmid. 2019. "Mars: Motion-Augmented Rgb Stream for Action Recognition." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA* 2019June: 7874–83. doi:10.1109/CVPR.2019.00807.

Dai, C., X. Liu, and J. Lai. 2020. Human Action Recognition Using Two-Stream Attention Based LSTM Networks. *Applied Soft Computing Journal* 86:105820. doi:10.1016/j.asoc.2019.105820.

Dai, X., X. Yuan, and X. Wei. 2021. Data Augmentation for Thermal Infrared Object Detection with Cascade Pyramid Generative Adversarial Network. *Appl Intell* 52: 967–981. https://doi.org/10.1007/s10489-021-02445-9

Dalal, N., B. Triggs, N. Dalal, and B. Triggs. 2005. "Histograms of Oriented Gradients for Human Detection To Cite This Version : Histograms of Oriented Gradients for Human Detection." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA*, 886–93. http://lear.inrialpes.fr .

Damen, D., H. Doughty, G. Maria Farinella, S. Fidler, A. Furnari, E. Kazakos, and D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray . 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11208 LNCS 753–71. doi: 10.1007/978-3-030-01225-0_44.

Deng, J., R. S. Wei Dong, L. Li-Jia, L. Kai, and L. Fei-Fei. 2009. *ImageNet: A Large-Scale Hierarchical Image Database*. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–55. doi: 10.1109/cvprw.2009.5206848.

Donahue, J., L. Anne Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. 2017. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (4):677–91. doi:10.1109/TPAMI.2016.2599174.

Duta, I. C., B. Ionescu, K. Aizawa, and N. Sebe. 2017. Spatio-Temporal VLAD Encoding for Human Action Recognition in Videos. *MultiMedia Modeling* 1 (November):226–37. doi:10.1007/978-3-319-51811-4.

Feichtenhofer, C., A. Pinz, and A. Zisserman. 2016. "Convolutional Two-Stream Network Fusion for Video Action Recognition." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016-Dec: 1933–41. doi:10.1109/CVPR.2016.213.

Gao, H., Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. 2017. "Densely Connected Convolutional Networks." *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, Hawaii, 2017-Janua: 2261–69. doi:10.1109/CVPR.2017.243.

Goyal, R., S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, and V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic . 2017. "The 'Something Something' Video Database for Learning and Evaluating Visual Common Sense." In *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017October:5843–51. doi:10.1109/ICCV.2017.622.

Hao, Y., C. Yuan, L. Bing, D. Yang, J. Xing, H. Weiming, and S. J. Maybank. 2018. Asymmetric 3D Convolutional Neural Networks for Action Recognition. *Pattern Recognition* 85:1–12. doi:10.1016/j.patcog.2018.07.028.

Hao, W., and Z. Zhang. 2019. Spatiotemporal Distilled Dense-Connectivity Network for Video Action Recognition. *Pattern Recognition* 92:13–24. doi:10.1016/j.patcog.2019.03.005.

Heilbron, F. C., and J. Carlos Niebles. 2014. "Collecting and Annotating Human Activities in Web Videos." *ICMR 2014 - Proceedings of the ACM International Conference on Multimedia Retrieval 2014*, Glasgow, Scotland, 377–84. doi:10.1145/2578726.2578775.

Heilbron, F. C., V. Escorcia, B. Ghanem, and J. Carlos Niebles. 2015. "ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, 07-12June: 961–70. doi:10.1109/CVPR.2015.7298698.

Herath, S., M. Harandi, and F. Porikli. 2017. Going Deeper into Action Recognition: A Survey. *Image and Vision Computing* 60:4–21. doi:10.1016/j.imavis.2017.01.010.

Htike, K. K., O. O. Khalifa, H. Adibah Mohd Ramli, and M. A. M. Abushariah. 2014. "Human Activity Recognition for Video Surveillance Using Sequences of Postures." *2014 3rd International Conference on E-Technologies and Networks for Development, ICeND 2014*, Beirut, Lebanon, 79–82. doi:10.1109/ICeND.2014.6991357.

Huang, C. Di, C. Yao Wang, and J. Ching Wang. 2016. "Human Action Recognition System for Elderly and Children Care Using Three Stream ConvNet." *Proceedings of 2015 International Conference on Orange Technologies, ICOT 2015*, HONG KONG, 5–9. doi:10.1109/ICOT.2015.7498476.

Huerta, E. A., A. Khan, E. Davis, C. Bushell, W. D. Gropp, D. S. Katz, and V. Kindratenko, S. Koric, William T. C. Kramer, B. McGinty, K. McHenry and A. Saxton . 2020. Convergence of Artificial Intelligence and High Performance Computing on NSF-Supported Cyberinfrastructure. *Journal of Big Data* 7 (1):1. doi:10.1186/s40537-020-00361-2.

Ilya, S., O. Vinyals, and Q. V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems* 4:3104–12.

Jegham, I., A. Ben Khalifa, I. Alouani, and M. Ali Mahjoub. 2020. Vision-Based Human Action Recognition: An Overview and Real World Challenges. *Forensic Science International: Digital Investigation* 32:200901. doi:10.1016/j.fsidi.2019.200901.

Jégou, H., F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. 2012. Aggregating Local Image Descriptors into Compact Codes. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* 34 (9):1704–16. doi:10.1109/TPAMI.2011.235.

Jhuang, H., J. Gall, S. Zuffi, C. Schmid, and M. J. Black. 2013. "Towards Understanding Action Recognition." In *Proceedings of the IEEE International Conference on Computer Vision*, Sydney, NSW, Australia, 3192–99. doi:10.1109/ICCV.2013.396.

Karpathy, A., G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei Fei 2014a. "Large-Scale Video Classification with Convolutional Neural Networks." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA. doi:10.1109/CVPR.2014.223.

Karpathy, A., G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. 2014b. "Large-Scale Video Classification with Convolutional Neural Networks." In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR, Columbus, OH, USA. doi:10.1007/978-981-15-7062-9_69.

Kay, W., J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, and F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman . 2017. The Kinetics Human Action Video Dataset. *ArXiv*

Kläser, A., M. Marszałek, and C. Schmid. 2008. "A Spatio-Temporal Descriptor Based on 3D-Gradients." In *BMVC 2008 - Proceedings of the British Machine Vision Conference 2008*, Leeds, UK. doi:10.5244/C.22.99.

Koohzadi, M., and N. Moghadam Charkari. 2017. Survey on Deep Learning Methods in Human Action Recognition. *IET Computer Vision* 11 (8):623–32. doi:10.1049/iet-cvi.2016.0355.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM* 60 (6):84–90. doi:10.1145/3065386.

Kuehne, H., H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: A Large Video Database for Human Motion Recognition, 2011 International Conference on Computer Vision, 2011, pp. 2556-2563, doi: 10.1109/ICCV.2011.6126543.

Kuehne, H., A. Arslan, and T. Serre. 2014. "The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities." In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 780–87. IEEE Computer Society. doi:10.1109/CVPR.2014.105.

Laptev, Ivan, M. Marszałek, C. Schmid, and B. Rozenfeld. 2008. "Learning Realistic Human Actions from Movies." *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, Anchorage, AK, USA, 0–7. doi:10.1109/CVPR.2008.4587756.

Lei, W., D. Q. Huynh, and P. Koniusz. 2020. A Comparative Review of Recent Kinect-Based Action Recognition Algorithms. *IEEE Transactions on Image Processing* 29:15–28. doi:10.1109/TIP.2019.2925285.

Lin, S., K. Jia, D. Yan Yeung, and B. E. Shi. 2015. "Human Action Recognition Using Factorized Spatio-Temporal Convolutional Networks." *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 2015 Inter: 4597–605. doi:10.1109/ICCV.2015.522.

Lin, S., K. Jia, K. Chen, D. Yan Yeung, B. E. Shi, and S. Savarese. 2017. "Lattice Long Short-Term Memory for Human Action Recognition." *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017October: 2166–75. doi:10.1109/ICCV.2017.236.

Liu, B., H. Cai, J. Zhaojie, and H. Liu. 2019. RGB-D Sensing Based Human Action and Interaction Analysis: A Survey. *Pattern Recognition* 94:1–12. doi:10.1016/j.patcog.2019.05.020.

Liu, Z., and H. Haifeng. 2019. Spatiotemporal Relation Networks for Video Action Recognition. *IEEE Access* 7:14969–76. doi:10.1109/ACCESS.2019.2894025.

Lowe, D. G. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60 (2):91–110. doi:10.1023/B:VISI.0000029664.99615.94.

Majumder, S., and N. Kehtarnavaz. 2021. Vision and Inertial Sensing Fusion for Human Action Recognition: A Review. *IEEE Sensors Journal* 21 (3):2454–67. doi:10.1109/JSEN.2020.3022326.

Michalis, V., C. Nikou, and I. A. Kakadiaris. 2015. A Review of Human Activity Recognition Methods. *Frontiers Robotics AI* 2 (NOV):1–28. doi:10.3389/frobt.2015.00028.

Minh Dang, L., K. Min, H. Wang, M. Jalil Piran, C. Hee Lee, and H. Moon. 2020. Sensor-Based and Vision-Based Human Activity Recognition: A Comprehensive Survey. *Pattern Recognition* 108:107561. doi:10.1016/j.patcog.2020.107561.

Monfort, M., A. Andonian, B. Zhou, K. Ramakrishnan, S. Adel Bargal, T. Yan, and L. Brown, Q. Fan, D. Gutfreund, C. Vondrick, and A. Oliva . 2020. Moments in Time Dataset: One Million Videos for Event Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42 (2):502–08. doi:10.1109/TPAMI.2019.2901464.

Naeem, B., Hajra, F. Murtaza, M. Haroon Yousaf, and S. A. Velastin. 2021. T-VLAD: Temporal Vector of Locally Aggregated Descriptor for Multiview Human Action Recognition. *Pattern Recognition Letters*. doi:10.1016/j.patrec.2021.04.023.

Najeera, P. M., P. D. Anu, and M. Sadiq. 2018. "An Intelligent Action Predictor from Video Using Deep Learning." *2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research, ICETIETR 2018*, Ernakulam, India, 2018–21. doi:10.1109/ICETIETR.2018.8529076.

Ng, J. Y. H., M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. 2015. "Beyond Short Snippets: Deep Networks for Video Classification." *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 07-12June: 4694–702. doi:10.1109/CVPR.2015.7299101.

Nweke, H. F., Y. Wah Teh, M. Ali Al-garadi, and U. Rita Alo. 2018. Deep Learning Algorithms for Human Activity Recognition Using Mobile and Wearable Sensor Networks: State of the Art and Research Challenges. *Expert Systems with Applications* 105:233–61. doi:10.1016/j.eswa.2018.03.056.

Özyer, T., A. Duygu Selin, and R. Alhajj. 2021. Human Action Recognition Approaches with Video Datasets—A Survey. *Knowledge-Based Systems* 222:106995. doi:10.1016/j.knosys.2021.106995.

Peng, X., L. Wang, X. Wang, and Y. Qiao. 2016. Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice. *Computer Vision and Image Understanding* 150 (September):109–25. doi:10.1016/j.cviu.2016.03.013.

Scovanner, P., S. Ali, and M. Shah. 2007. "A 3-Dimensional Sift Descriptor and Its Application to Action Recognition." In *Proceedings of the ACM International Multimedia Conference and Exhibition*, Augsburg,Germany, 357–60. doi:10.1145/1291233.1291311.

Serpush, F., and M. Rezaei. 2021. Complex Human Action Recognition Using a Hierarchical Feature Reduction and Deep Learning-Based Method. *SN Computer Science* 2 (2):1–15. doi:10.1007/s42979-021-00484-0.

Sheng, Y., L. Xie, L. Liu, and D. Xia. 2020. Learning Long-Term Temporal Features with Deep Neural Networks for Human Action Recognition. *IEEE Access* 8:1840–50. doi:10.1109/ACCESS.2019.2962284.

Shuiwang, J., X. Wei, M. Yang, and Y. Kai. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (1):221–31. doi:10.1109/TPAMI.2012.59.

Sigurdsson, G. A., G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. 2016. "Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding." http://allenai.org/plato/charades/.

Simonyan, K., and A. Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. *Advances in Neural Information Processing Systems* 1 (January):568–76.

Simonyan, K., and A. Zisserman. 2015a. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, San Diego, CA, 1–14.

Simonyan, K., and A. Zisserman. 2015b. "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION." http://www.robots.ox.ac.uk/ .

Singh, T., and D. Kumar Vishwakarma. 2019. Video Benchmarks of Human Action Datasets: A Review. *Artificial Intelligence Review* 52 (2):1107–54. doi:10.1007/s10462-018-9651-1.

Soomro, K., A. Roshan Zamir, and M. Shah. 2012. "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild". November. http://arxiv.org/abs/1212.0402.

Srivastava, N., E. Mansimov, and R. Salakhutdinov. 2015. "Unsupervised Learning of Video Representations Using LSTMs." *32nd International Conference on Machine Learning, ICML 2015*, Lille, France, 1: 843–52.

Sun, B., W. Yong, K. Zhao, H. Jun, Y. Lejun, H. Yan, and A. Luo. 2021. Student Class Behavior Dataset: A Video Dataset for Recognizing, Detecting, and Captioning Students' Behaviors in Classroom Scenes. *Neural Computing & Applications* 0123456789. doi:10.1007/s00521-020-05587-y.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. "Going Deeper with Convolutions." In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA*, 07-12June:1–9. doi:10.1109/CVPR.2015.7298594.

Taylor, G. W., R. Fergus, Y. LeCun, and C. Bregler. 2010. Convolutional Learning of Spatio-Temporal Features. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 6316 LNCS (PART 6:140–53. doi:10.1007/978-3-642-15567-3_11.

Thameri, M., A. Kammoun, K. Abed-Meraim, and A. Belouchrani. 2011. "Fast Principal Component Analysis and Data Whitening Algorithms." *7th International Workshop on Systems, Signal Processing and Their Applications, WoSSPA 2011, Tipaza, Algeria*, 139–42. doi:10.1109/WOSSPA.2011.5931434.

Tran, D., L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. 2015. "Learning Spatiotemporal Features with 3D Convolutional Networks." *Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile* 2015 Inter: 4489–97. doi:10.1109/ICCV.2015.510.

Varol, G., I. Laptev, and C. Schmid. 2018. Long-Term Temporal Convolutions for Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (6):1510–17. doi:10.1109/TPAMI.2017.2712608.

Verma, K. K., H. L. M. Brij Mohan Singh, and P. Chauhan. 2020. Two-Stage Human Activity Recognition Using 2D-ConvNet. *International Journal of Interactive Multimedia and Artificial Intelligence* 6 (2):11. doi:10.9781/ijimai.2020.04.002.

Verma, K. K., and B. Mohan Singh. 2021. Deep Multi-Model Fusion for Human Activity Recognition Using Evolutionary Algorithms. *International Journal of Interactive Multimedia and Artificial Intelligence* 7 (2):44–58. doi:10.9781/ijimai.2021.08.008.

Verma, K. K., B. Mohan Singh, and A. Dixit. 2022. A Review of Supervised and Unsupervised Machine Learning Techniques for Suspicious Behavior Recognition in Intelligent Surveillance System. *International Journal of Information Technology (Singapore)* 14 (1):397–410. doi:10.1007/s41870-019-00364-0.

Wan, S., Q. Lianyong, X. Xiaolong, C. Tong, and G. Zonghua. 2020a. Deep Learning Models for Real-Time Human Activity Recognition with Smartphones. *Mobile Networks and Applications* 25 (2):743–55. doi:10.1007/s11036-019-01445-x.

Wan, Y., Y. Zujun, Y. Wang, and L. Xingxin. 2020b. Action Recognition Based on Two-Stream Convolutional Networks with Long-Short-Term Spatiotemporal Features. *IEEE Access* 8:85284–93. doi:10.1109/ACCESS.2020.2993227.

Wang, H., A. Klaser, C. Schmid, and C.-L. Liu. 2011. Action Recognition by Dense Trajectories. *IEEE Access*. doi:10.16182/j.1004731x.joss.201709023.

Wang, H., and C. Schmid. 2013. "Action Recognition with Improved Trajectories." *Proceedings of the IEEE International Conference on Computer Vision*, 3551–58. doi:10.1109/ICCV.2013.441.

Wang, H., A. Kläser, C. Schmid, and C. Lin Liu. 2013a. Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *International Journal of Computer Vision* 103 (1):60–79. doi:10.1007/s11263-012-0594-8.

Wang, H., C. Schmid, H. Wang, C. Schmid, A. Recognition, T. Iccv, H. Wang, and C. Schmid. 2013b. "Action Recognition with Improved Trajectories To Cite This Version : HAL Id : Hal-00873267 Action Recognition with Improved Trajectories." *ICCV - IEEE International Conference on Computer Vision, Sydney, NSW, Australia* December: 3551–58.

Wang, L., Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. van Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9912 LNCS: 20–36. doi:10.1007/978-3-319-46484-8_2.

Wang, X., Z. Miao, R. Zhang, and S. Hao. 2019. "I3D-LSTM: A New Model for Human Action Recognition." *IOP Conference Series: Materials Science and Engineering* 569 (3). doi:10.1088/1757-899X/569/3/032035.

Wang, S., Y. Liu, J. Wang, S. Gao, and W. Yang. 2021. A Moving Track Data-Based Method for Gathering Behavior Prediction at Early Stage. *Applied Intelligence* 51 (11):8498–518. doi:10.1007/s10489-021-02244-2.

Wenjin, T., Z. Hao Lai, M. C. Leu, and Z. Yin. 2018. Worker Activity Recognition in Smart Manufacturing Using IMU and SEMG Signals with Convolutional Neural Networks. *Procedia Manufacturing* 26:1159–66. Elsevier B.V. doi:10.1016/j.promfg.2018.07.152.

Zawar, H., Q. Z. Sheng, and W. Emma Zhang. 2020. A Review and Categorization of Techniques on Device-Free Human Activity Recognition. *Journal of Network and Computer Applications* 167:102738. December 2019. doi:10.1016/j.jnca.2020.102738.

Zhang, S., Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li. 2017. A Review on Human Activity Recognition Using Vision-Based Method. *Journal of Healthcare Engineering* 2017:1–31. doi:10.1155/2017/3090343.

Zhang, H.-B., Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, D. Ji-Xiang, and D.-S. Chen. 2019. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. *Mpdi*. doi:10.3390/s19051005.

Zhao, H., and X. Jin. 2020. "Human Action Recognition Based on Improved Fusion Attention CNN and RNN." *Proceedings - 2020 5th International Conference on Computational Intelligence and Applications, ICCIA 2020, Beijing, China*, 108–12. doi:10.1109/ICCIA49625.2020.00028.

Zheng, H., and X.-M. Zhang. 2020. "A Cross-Modal Learning Approach for Recognizing Human Actions." *IEEE Systems Journal*, 1–9. doi:10.1109/jsyst.2020.3001680.

Zhenyang, L., K. Gavrilyuk, E. Gavves, M. Jain, and C. G. M. Snoek. 2018. VideoLSTM Convolves, Attends and Flows for Action Recognition. *Computer Vision and Image Understanding* 166:41–50. doi:10.1016/j.cviu.2017.10.011.

Zhu, F., L. Shao, J. Xie, and Y. Fang. 2016. From Handcrafted to Learned Representations for Human Action Recognition: A Survey. *Image and Vision Computing* 55:42–52. doi:10.1016/j.imavis.2016.06.007.