Zhang and Wang
EURASIP Journal on Image and Video Processing (2025) 2025:15
https://doi.org/10.1186/s13640-025-00677-0

EURASIP Journal on Image and Video Processing

REVIEW Open Access

A comprehensive survey on RGB-D-based human action recognition: algorithms, datasets, and popular applications

Yumin Zhang[1] and Yanyong Wang[1*]

*Correspondence:
wangyanyongdl@126.com

[1] Department of Weapon Control System, China North Vehicle Research Institute, Dahuichang East Street, 100072 Beijing, Beijing, China

Abstract

Due to the rapid advances in computer vision and deep learning, human action recognition has become one of the most important representative tasks for video understanding. Especially for human action recognition based on RGB-D data, a promising research direction, there has been a number of researchers to work on. In particular, convolutional neural networks (CNNs) are capable of image classification tasks, recurrent neural networks (RNNs) are skilled in sequence-based problems, and Transformer is good at global modeling. In this survey, we introduce a number of algorithms based on CNNs, RNNs and Transformer for RGB-D based human action recognition, which could be categorized into four parts: RGB-based, depth-based, skeleton-based and RGB-D based. As a survey focusing on the RGB-D based human action recognition, we thoroughly represent the algorithms, datasets and popular applications for it. What's more, we give some possible future research directions for this field in the last part.

Keywords: Human action recognition, Convolutional neural networks, RGB-D data, Transformer, Recurrent neural networks

1 Introduction

Human action recognition (HAR) has recently garnered significant attention among computer vision researchers, finding applications in diverse fields such as robot vision, multimedia content retrieval, video surveillance, and motion tracking systems. Particularly, the rapid improvement of low-cost sensors, including Microsoft Kinect [1], Intel RealSense [2] and Orbbec [3] has stimulated additional research endeavors in the domain of action recognition. Each of the modalities mentioned before possesses distinct characteristics that facilitate addressing challenges associated with action data and offer promising avenues for computer vision researchers to explore vision data from diverse perspectives.

Early research on HAR was dominated by the research in the still images or videos [4], locating people in a video frame both in the spatial dimension and the temporal dimension with bounding boxes, temporal extent and a spatial–temporal cuboid which contains a special action. Action recognition continues to overcome significant

challenges, primarily due to the issues arising from background clutter, partial occlusion, variations in viewpoint and lighting, differences in execution rates, and biometric variability.Having a precise understanding of human actions in still images is a tough task, so many researchers change their direction from still images to videos. Depth data significantly contribute to HAR by providing complementary 3D structural information and enhancing robustness. It offers 3D coordinates of skeletal joints (e.g., Kinect's 20 joints), crucial for modeling spatio-temporal dynamics in complex networks [5]. Insensitive to illumination and occlusion, it outperforms RGB in challenging scenarios, as shown in Khaire et al. [6]. When fused with RGB and skeleton modalities, depth data boost recognition via feature/decision-level fusion: for example, dynamic images from depth and RGB enhance discriminative representations in transformers [7], while skeleton-guided depth ROI generation focuses on action-critical regions [8]. Depth data thus bridge modal limitations, enabling robust, accurate HAR and inspiring future multimodal integrations.

Several survey papers have reviewed the research on video action recognition with RGB-D data.Specifically, Chen et al.[9] concentrated on depth sensors, the preprocessing of depth data, depth-based methods for action recognition, and relevant datasets. Their work provided a comprehensive overview of methodologies that combined depth and skeleton modalities for various tasks, such as action recognition, estimation of head or hand poses, facial feature detection, and gesture recognition. Aggarwal and Xia[10] gave a summary about five categories of representations, which were based on 3D silhouettes, skeletal joints or body part locations, local spatial–temporal features, scene flow features, and local occupancy features. Besides, Cheng et al. [11] concentrated on RGB-D-based datasets for hand gesture recognition and provided a comprehensive summary of the corresponding methodologies from three distinct perspectives: static hand gesture recognition, hand trajectory-based gesture recognition, and continuous hand gesture recognition. What's more, Escalera et al. [12] conducted a thorough review of the challenges and methodologies associated with gesture recognition with multimodal data. Furthermore, several surveys within this domain have specifically emphasized the available datasets pertinent to RGB-D research.For instance, Zhang et al. [13] detailed the benchmark RGB-D datasets available for action recognition. These datasets encompassed 27 single-view datasets, 10 multi-view datasets, and 7 multi-person datasets. And Han et al. [14] thoroughly gave a review about skeleton-based representation and the algorithms for action recognition. After that, Yao et al. [15]provided a review of convolutional neural network-based action recognition and indicated the limitations and directions for CNN-based action recognition. The work of Sun et al. [16] was the first survey paper that thoroughly review the HAR algorithms from the perspective of all kinds of data modalities, including RGB, depth, skeleton, infrared sequence, point cloud, event stream, audio, acceleration, radar and WiFi. However, three critical limitations persist in these prior efforts. First of all, existing reviews [9, 10] and [14] focus predominantly on conventional RGB-D or skeleton data, overlooking emerging sensing modalities like WiFi-CSI and mmWave radar that enable robust perception in occlusion-rich scenarios; secondly, the methodological analyses in [11] and [15] primarily cover pre-2020 CNN architectures, neglecting transformative paradigms such as vision transformers (e.g., VideoMAE) and neural radiance fields

(NeRFs); finally, while dataset benchmarks[13] catalog lab-controlled recordings, fewer than 15% of evaluated methods [12] and [16] address challenges like dynamic illumination changes or multi-person interactions in unconstrained environments.

As for this review, we provide a comprehensive review of RGB-D-based human action recognition, including RGB, depth, skeleton and RGB-D. Some relevant surveys, such as those in [17–20], study from a single modality perspective or compare the characteristics of different datasets. To the best of our knowledge, there is no review before which concentrates on the methods fusion with other fields in computer vision and the applications of action recognition for RGB-D based action recognition. In the field of video action recognition, classification techniques extract distinctive features from each modality and employ sophisticated computer vision methodologies.

A novel contribution of this review is the concentration on RGB-D data-based action recognition combined with other fields in computer vision. Moreover, this paper distinguishes itself from other studies by the following contributions:

1. Review of the applications in the combination of action recognition with other subfields in computer vision in order that readers could have a comprehensive overview of the advanced techniques.
2. Analysis of the more recent and advanced deep learning algorithms for HAR, and hence provide the readers with the state-of-the-art approaches.
3. Review of the multi-modality-based HAR algorithms, including RGB, depth, skeleton and RGB-D.
4. Discussion of the challenges of data fusion and action recognition and potential future research directions.

The reminder of this paper is organized as follows. Section 2 discusses the state-of-the-art algorithms in different modalities, including RGB, depth, skeleton and RGB-D. Section 3 introduces the benchmark datasets for human action recognition and data acquisition. Section 4 reviews the algorithms combined with other subfields in computer vision based on human video action recognition. Section 5 outlines the applications of human action recognition in different areas. And finally Section 6 gives some possible promising research directions and Section 7 concludes the whole review.

2 Different modalities algorithms

In this section, we comprehensively review the algorithms using RGB-D data, including RGB data [13], depth data [21, 22], skeleton data[23] and the combination with these data modalities[24]. The following subsections will introduce RGB-, depth-, skeleton- and RGB-D based algorithms.

2.1 RGB-based

RGB constitutes a pivotal channel within RGB-D data, characterized by attributes such as shape, color, and texture, which encompass a wealth of distinctive features. These properties further render it highly efficacious for the direct utilization of networks. Although most of the surveyed algorithms for this section are not directly applied into RGB-D based datasets, it is also proper to use these methods to the RGB modality of

RGB-D datasets.Table 1 and Table 1(Continued) give a comprehensive summary for the RGB-based HAR algorithms.

2.1.1 Hand-crafted methods

Before the existence of DeepVideo[1], people used to study video action recognition with hand-crafted features, such as using dense trajectories[25], using improved trajectories[26], using stacked fisher vectors[27], Multi-skip Feature Stacking(MIFS)[28] .What's more, Improved Dense Trajectories (IDT) is a typical method during the stage of hand-crafted features.

IDT took camera motion into account. They matched feature points between frames using SURF[29] descriptors and dense optical flow. Since human motion is different from camera motion and generate inconsistent matches, IDT used a human detector to improve the estimation. Besides, this estimation was used to cancel out camera motion from the optical flow. Before applying CNN to video action, IDT was the state of art method.

2.1.2 2D-CNN-based methods

Videos could naturally be decomposed into spatial part and temporal part. The former one conveys the information on objects of the real world in the video, and the latter one, in the form of motion across the frames, exhibits the movement of the camera recording the position change of the objects in the video. As a result, using CNNs for both spatial dimension and temporal dimension could improve the performance of CNN-based video action recognition. There are four kinds of approaches to encode spatial–temporal information.

The first approach was established by Two-Stream Networks[30], which created a second path to get knowledge of the temporal information through training a convolutional neural networks with optical flow stream. They divided the video recognition architecture into spatial stream and temporal stream, as shown in Fig. 1. Since the spatial

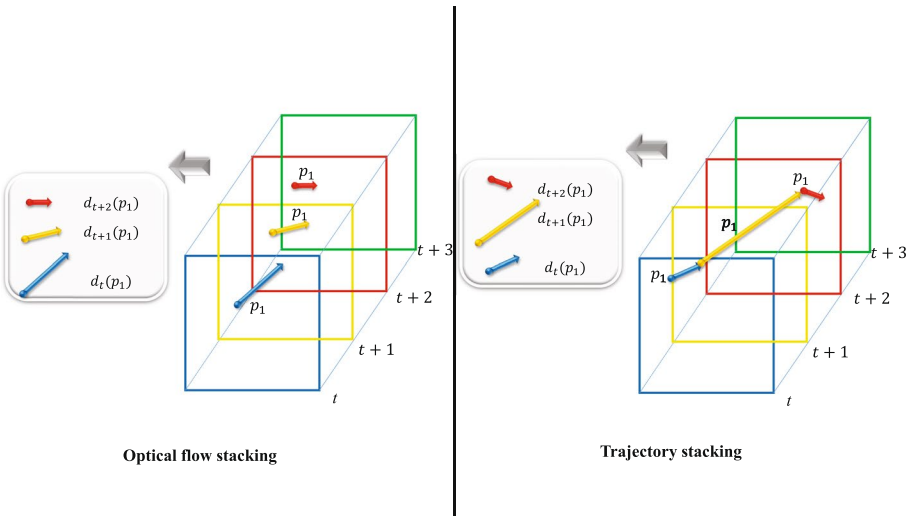

Fig. 1 ConvNet input derivation from optical flow stacking and trajectory

ConvNet ought to do is more like the task of image classification, Simonyan et al.[30] built upon the large-scale image classification methods[31] and pre-trained the network on a large image classification dataset, such as ImageNet dataset.

And the second approach is to use deeper network architecture. Wang et al. [32] discovered that merely employing deeper neural networks did not necessarily translate into superior performance. This observation may be attributed to the phenomenon of overfitting, particularly when dealing with small-sized video datasets as reported in [33]. As a result, Wang et al. [34] presented a suite of best practices aimed at mitigating overfitting in deeper networks. These practices covered cross-modality initialization, synchronized batch normalization, corner and multi-scale cropping for data augmentation, and the application of a substantial dropout ratio. With these good practices, it was capable of training the two-stream network proposed by Wang et al. [34] with the VGG-16 model [35] and outperforms two-stream network [30] by a large margin.

As for the third approach, it is called two-stream fusion. First of all, we would like to introduce three major fusions, early fusion, late fusion and slow fusion. Across temporal domain, fusion can be done early in the network by modifying the first layer convolutional filters to extend in time, which is called early fusion. Or it could be done late by placing two separate single-frame networks some distance in time apart and fusing their outputs later in the processing. As for slow fusion, it seems to be like the mixture of early fusion and late fusion, which slowly fuses temporal information throughout the network such that higher layers get access to progressively more global information in both spatial and temporal dimensions. Feichtenhofer et al.[36] shows that early fusion is beneficial for both streams to learn richer features and leads to improved performance over late fusion. Simonyan et.al [30] and Wang et.al [34] showed that late fusion could make the two-stream ConvNet work better. Late fusion takes place at the decision level, where the predictions or scores from different classifiers are combined to produce the final output. Fig. 2 precisely shows the mechanism of three fusions. Recent years, Yang et al. [37] proposed the Two-Stream Densenet-3D (TS-D3D) model for HAR, combining 3D DenseNet with a two-stream approach. It includes an RGB pathway for spatial features and an RGB DIFF pathway for temporal features, with modified convolution kernels in the latter to better capture temporal correlations and reduce computation. A transition layer fuses features from both pathways. The model shows strong performance on standard HAR datasets.

For the last approach, they are called segment-based methods. Due to the great success of optical flow, two-stream networks are able to reason about short-term motion information between frames. However, it is not capable of capturing long-term temporal information. In order to overcome this weakness, Wang et al. [38] proposed P-CNN based on the idea of long-range temporal structure modeling. It combines a sparse temporal sampling strategy and video-level supervision to enable efficient and effective learning with the whole action video. As shown in Fig. 3, Wang et al. [39] proposed TSN, which first divides a whole video into several segments, and the segments distribute uniformly along the temporal dimension. Then TSN randomly selects a single video frame, which is also called a short snippet in the original paper. The class scores of different snippets are fused by a segmental

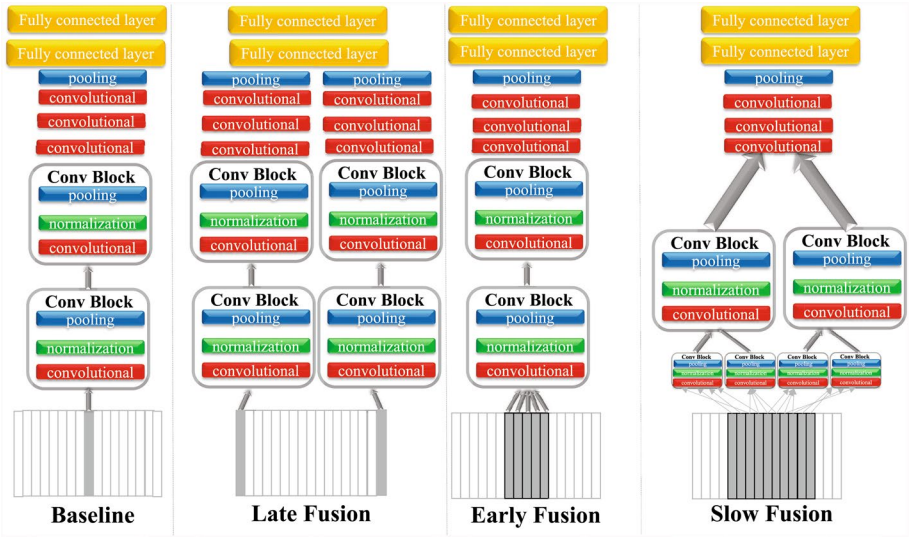

Fig. 2 The architecture of different fusions (the second layer of slow fusion in conv block same to other fusions)

consensus function to obtain segmental consensus, which is a video-level prediction. Subsequently, predictions from all modalities are combined to generate the final prediction. ConvNets on all snippets share the same parameters. Due to the proper design, TSN is able to model long-range temporal structure because the model is capable of seeing the content from the whole video. Besides, this sparse sampling strategy lowers the training cost over long video sequences but preserves relevant information. Furthermore, Umran et al. [40] proposed an innovative method to improve HAR in video surveillance, overcoming manual monitoring limitations. They use an unsupervised I3D [41] to identify regions of interest (ROI) within videos, focusing on areas with object movement. These ROI frames are then compiled into a new video sequence. Temporal Segment Networks (TSN) classify and recognize human actions from these ROI videos, enhancing both efficiency and accuracy.

2.1.3 3D-CNN-based methods

The breakthrough work for using 3D CNNS for action recognition is [42]. As it was a seminal work, it was not deep enough to show its lasting potential. Tram et al. [43] extended [42] to C3D. C3D follows the modular design of [35], and it was regarded as a 3D version of VGG16 network. Although its performance on standard benchmark datasets is not satisfying, it shows a strong generalization capability and could be used as a tool for many other video tasks [44]. However, 3D networks are not easy to optimize. In order to train a 3D convolutional filter well, researchers need a large-scale dataset with various video content and action categories. Sports 1M [1] puts this area forward and people could train a deep 3D network with the help of it. Unfortunately, the training of C3D takes weeks to converge. Though C3D was really popular, most users just set C3D as a feature extractor instead of using it to modify the network. This is partially the reason why two-stream networks based on 2D

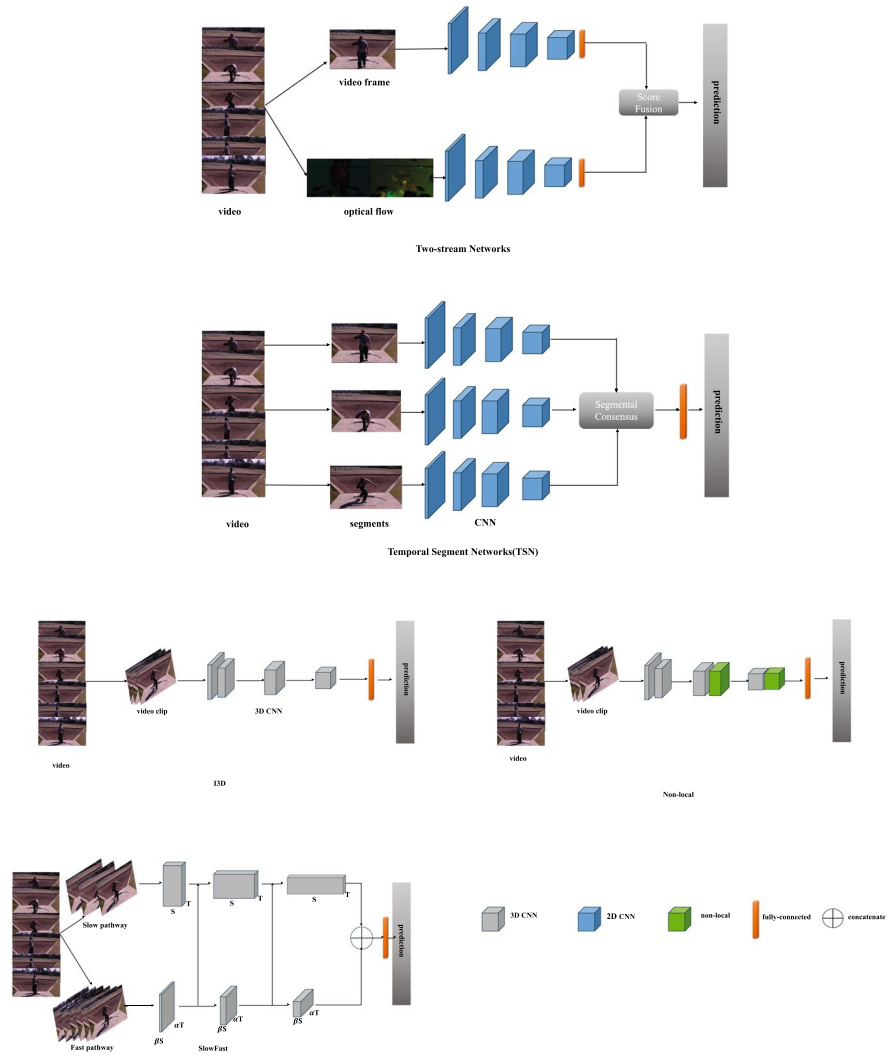

Fig. 3 The architecture of five papers: two-stream networks, TSN, I3D, Non-local and SlowFast

CNNs dominated the video action recognition domain from year 2014 to 2017. In this section, we would propose four approaches for 3DCNN-based methods.

From 2D CNNs to 3D CNNs2D CNNs take the benefits of pre-training brought by the large-scale image datasets such as ImageNet [45] and Places205 [46], both of which are far larger and more diverse than today's biggest video datasets,enabling effective feature learning for visual tasks.This tough situation was changed by the existence of I3D [41] proposed by Carreira et al. in 2017. As shown in Figure 3,the I3D model processes a video clip by passing it through a stack of 3D convolutional layers. A video clip is a sequence of video frames, often 16 or 32 frames. The major contributions of I3D are: (1) The authors adapted mature image classification architectures to use for 3D CNN; (2) for model weights, the authors adopted a method developed for initializing optical flow networks in [34] to expand the ImageNet pre-trained 2D model weights to their counterparts in the 3D model. Therefore, I3D bypassed the dilemma that 3D CNNs

have to be trained from scratch. From then on, papers following I3D needed to test the performance of their network on Kinetics, or other large-scale benchmark datasets, which pushed video action recognition forward to the next level. Inspired by I3D, ResNet3D [47] directly took 2D ResNet [48] and replaced all the 2D convolutional filters with 3D kernels. They thought that by using deep 3D CNNs together with large-scale datasets one can exploit the success of 2D CNNs on ImageNet. Inspired by ResNext[49], Chen et al. [50] presented SENet, a multi-fiber architecture that slices a complex neural network into an ensemble of lightweight networks that run through the network, which is beneficial for information flow between fibers and reduces the computational cost at the same time. Furthermore, motivated by SENet [50], Diba et al. [51] proposed STCNet, integrating channel-wise information inside a 3D block to capture both spatial-channels and temporal-channels correlation information throughout the whole network. In the next few years, 3D CNNs advanced quickly and became top performers on almost every benchmark dataset.

Integrating 2D and 3D CNNs To reduce the complexity of 3D network training, P3D [52] and R(2+1)D [53] explore the idea of 3D factorization. To be specific, a 3D kernel (e.g., 3×3×3) could be factorized to two separate operations, a 2D spatial convolution and a 1D temporal convolution(e.g., 1×3×3 ,3×3×3).The differences between P3D and R(2+1)D are how they arrange the two factorized operations and how they formulate each residual block. Trajectory convolution [54] follows this idea but uses deformable convolution for the temporal component to better cope with motion. Another way of simplifying 3D CNNs is to mix 2D and 3D convolutions in the same network. Zhou et al. [55] proposed MiCTNet, which integrates 2D and 3D CNNs to generate deeper and more informative feature maps, reducing training complexity in each round of spatio-temporal fusion. Wang et al. [56] proposed ARTNet, introducing an appearance-and-relation network by using a brand-new building block, which decouples the spatio-temporal learning module into an appearance branch for temporal modeling. Xie et al. [57] proposed S3D, which aggregates the strengths of prior methods through a hierarchical design. It replaces the 3D convolutions at the bottom of the network with 2D kernels, while retaining 3D convolutions in higher layers for temporal modeling. This structure treats the remaining 3D kernels as P3D and R(2+1)D do, making the model size smaller and the training complexity less. Similar to P3D and R(2+1)D, this design applies 3D kernel factorization only in deeper network stages. This selective spatio-temporal processing achieves a balance between accuracy (Kinetics: +1.2% vs. pure 2D) and efficiency (GPU memory: -28% vs. pure 3D). Recently, Men et al. [58] proposed an architecture based on 3D CNN and SlowFast for offline recognition of gestures in open surgery. Their approach utilized R3D and R(2+1)D for feature extraction. On a self-constructed open surgery dataset, their method achieved notable performance in surgical gesture recognition, with an accuracy of 90.4%, precision of 90.5%, and recall of 90.0%. Furthermore, Chang et al. [59] employed R(2+1)D for spatio-temporal feature extraction. This algorithm enables real-time analysis of athletes' movements during competitions, assisting coaches and team members in better understanding game dynamics, optimizing training methods, and formulating more effective competition strategies. Through this approach, it not only enhances the precision and efficiency of training but also provides a scientific basis for improving the team's competitive

performance. Moreover, Murugan et al. [60] proposed an R(2+1)D CNN architecture to improve human action recognition from videos. This method uses a factorized CNN with an optimized residual model to reduce parameter layers, addressing issues of vanishing gradients and computational complexity.

Long-range temporal modeling In order to achieve long-range temporal modeling, Varol et al. [61] proposed LTC and evaluated long-term temporal convolutions over a large number of video frames. However, due to the limitation by the memory of GPU, they have to lower the input resolution to take in more frames. Then, Diba et al. [62] proposed T3D, taking the benefits from DenseNet [63] and extending it with 3D convolutional filters and pooling kernels, thereby preserving the raw temporal information, which enhances prediction robustness against noisy inputs. After that, Wang et al. proposed Non-local [64], which is a generic operation modified from self-attention [65]. It can be applied in a huge number of computer vision tasks with a plug-and-play manner. As shown in Figure 3, they used a space–time non-local module after later residual blocks to capture the long-range dependency in both space and temporal domain, which computed the response at a position as a weighted sum of the features at all positions. And they achieved better performance on both Kinetics and Charades datasets even without any bells and whistles. In addition, Ha et al. [67]proposed DCapsNet, a top-heavy capsule network incorporating a spatio-temporal Non-local mechanism, which integrates 3D CNNs and leverages the Non-local mechanism to effectively recognize human actions by exploiting spatio-temporal contextual information in videos. Moreover, Elmadany et al.[68] proposed a Nonlocal Multi-Fiber Network (Nl-MFN) for HAR. By integrating the Non-local mechanism with a multi-fiber architecture, this model can more effectively capture long-range dependencies and complex spatio-temporal dynamics in videos, thereby achieving effective action recognition. What's more, Dong et al. [69]proposed a Mixed Time-Asymmetric Convolutional Neural Network (MTA CNN) that leverages time-asymmetric convolutions to extract non-local temporal features and employs conventional convolutions to capture local temporal features. By integrating both local and non-local temporal features, the MTA CNN achieves higher action recognition accuracy while maintaining a lightweight network structure and fast processing speed. Notably, the algorithm designs a temporal feature fusion method as an alternative to the common global average pooling, aiming to obtain higher-dimensional feature vectors and retain more information.

Enhancing 3D efficiency. Inspired by the development in efficient 2D networks, researchers started to apply channel-wise separable convolution in video classification [70, 71]. Then, Du et al.[71] introduced CSN, factorizing 3D convolutions by separating channel interactions and spatio-temporal interactions, and able to obtain state-of-the-art performance while being 2 to 3 times faster than the previous best approaches. All the approaches mentioned above are also related to multi-fiber networks [72] since they are all encouraged by group convolution. Later, Feichtenhofer et al. [73] introduced SlowFast, a powerful network for video action recognition with a slow pathway and a fast pathway. The former one operates at low frame rate to capture spatial semantics, while the latter one operates at high frame rate so that it could capture motion at fine temporal resolution. The fast pathway could be made very lightweight by reducing its channel capacity, yet could learn useful temporal information for video action recognition.

This network is encouraged by the biological parvocellular cells and magnocellular cells. Parvocellular cells could provide fine spatial detail and color, but lower temporal changes, while magnocellular cells always operate at high temporal frequency and are responsive to fast temporal changes, but not sensitive to spatial detail or color. And the fast pathway is analogous to magnocellular cells, while the slow pathway is analogous to parvocellular cells. This model has two pathways, but it is different from the two-stream networks, since the two pathways are designed to model different temporal speeds, rather than spatial and temporal modeling. Then, there is another paper which uses multiple pathways to balance the accuracy and efficiency[74]. Recent years, inspired by SlowFast [73], Kowshilk et al. [75] addressed the challenges of limited data volume and low resource availability in surveillance datasets by employing transfer learning and fine-tuning the I3D model along with SlowFast [73]. This approach enables the automatic extraction of features from surveillance videos in the SPHAR dataset and classifies them into the corresponding action categories.

2.1.4 RNN-based methods

Videos consist of sequentially ordered frames, thereby encoding temporal information in their sequential progression. As a result, researchers have explored Recurrent Neural Networks (RNNs) for temporal modeling inside a video.

The evolution of RNN-based HAR has progressed from foundational CNN-LSTM frameworks, such as LRCN [76] and Beyond-Short-Snippets [77], which aggregated frame-level features into video-level predictions through two-stream architectures and late fusion, to advanced innovations that address limitations in temporal modeling and computational efficiency. Subsequent studies, such as hierarchical multi-granularity LSTM network [78] for multi-scale feature aggregation. Li et al. [79] proposed Video-LSTM, including a correlation-based spatial attention mechanism and a lightweight motion-based attention mechanism. Video-LSTM not only shows improved results on action recognition, but also clarifies how the learned attention can be used for action localization by just relying on the action class label. Moreover, Sun et al. [80] proposed Lattice-LSTM, extending LSTM by learning independent hidden state transitions of memory cells for individual spatial locations, so that it could concretely model long-term and complex motions. Shi et al. [81] proposed ShuttleNet, representing a concurrent research effort that focuses on incorporating both feed-forward and feedback connections within a RNN framework to learn long-term dependencies. Several variants of the CNN-LSTM architecture have been proposed, such as bidirectional LSTM [82] (capturing bidirectional temporal dependencies), two-stream LSTM [83] (a dual-path design for spatial–temporal feature extraction), three-stream LSTM [84] (extending the dual-path concept to three modalities), and hierarchical CNN-LSTM [85] (a multi-level fusion framework).YIn et al. [86] proposed LSTM CrossRWKV, introducing RWKV into the video domain with a framework named LSTM CrossRWKV(LCR) for spatio-temporal representation learning. The LCR framework features a Cross RWKV gate to enhance interaction between current frame edge information and past features, improving focus on the subject and aggregating inter-frame features globally over time. Moreover, Zhu et al. [87] proposed a FAST-GRU architecture to aggregate segment-level features from both high-cost and low-cost backbone networks. The FAST-GRU

framework integrates high-quality representations from complex models for detailed action capture and uses lightweight models to cover scene changes, ensuring efficient video understanding with lower computational costs.

2.1.5 Transformer-based methods

Transformer [65] based solely on attention mechanisms, dispenses with recurrence and convolutions entirely. While this architecture has become the de facto standard for natural language processing tasks, its application to computer vision remain limited. In computer vision, attention is either applied combined with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. Now, for HAR, there are some approaches with Transformer and they achieved great performance on some widely used HAR datasets.There are two parts of Transformer-based methods.

Transformer only for human action recognition. A convolution-free approach to video classification can be built exclusively on self-attention over space and time. Bertasius et al. [88] proposed TimeSformer, applying the standard Transformer architecture to HAR by enabling spatio-temporal feature learning directly from a sequence of frame-level patches. They found that when temporal attention and spatial attention are separately applied with each block, it would lead to the best video classification accuracy among the design choices considered. TimeSformer achieved a top-1 accuracy of 82.2% on Kinetics-600, surpassing all existing methods at the time of its publication in 2021.In addition, compared to 3D convolutional networks, TimeSformer is faster to train and it achieved dramatically higher test efficiency(at a small drop in accuracy), and it could also be applied to much longer video clips, even over one minute long. Girdhar et al. [89] introduced VTn, another typical method that only use Transformer for HAR. The authors were inspired by the developing progress in vision transformers, and they ditched the standard approach in HAR that relied on 3D ConvNets and introduced a method that classified actions by attending to the entire video sequence information. In terms of wall runtime, it trained 16.1× faster and ran 5.1× faster during inference while maintaining competitive accuracy compared to other state-of-the-art methods. It enabled the whole video analysis, via a single end-to-end pass, while still requiring 1.5× fewer GFLOPs.

Transformer modified from CNNs for human action recognition The vision community is witnessing a modeling shift from CNNs to Transformers, where pure Transformer architectures have attained top accuracy on the major video recognition benchmarks. All these video networks are all built on Transformer layers that globally connect patches across the spatial and temporal dimensions. However, when we adopt self-attention [65] to model pixel-level long-range dependency for visual recognition tasks, especially the HAR tasks, the performance of the network can be better. Cao et al. [90] proposed GCNet, extending Non-local and creating a simplified network based on a query-independent formulation, which maintains the accuracy of Non-local but with obviously less computation. They unified it into a three-step general framework for global context modeling. Within the general framework, they designed the global context (GC) block, which is lightweight and able to effectively model the global context. Moreover, Khazaei et al. [91] introduced CDFL, also extending the Non-local method

and designing a simplified network based on a query-independent formula. This network maintained the accuracy of Non-local but significantly reduces the computational load. They unified these approaches into a three-step general framework aimed at global context modeling. Within this general framework, a better instantiated module called the Global Context (GC) block was designed. It was relatively lightweight and capable of efficiently modeling the global context.Liu et al.[92] proposed Video Swin Transformer, presenting a pure-transformer backbone architecture for HAR, similar to the convolutional models whose backbone architectures for video are adapted from those for images simply by extending the modeling by the temporal axis. This work was through a spatio-temporal adaptation of Swin Transformer [93], a general-purpose vision backbone for image classification and it incorporates bias for spatial locality, as well as for hierarchy and translation invariance. Video Swin Transformer follows this hierarchy structure and expand the applying field from the only spatial domain to the spatio-temporal domain. Similar to the receptive field in CNNs, this model has shifted window, which is originated from Swin Transformer, could be used to process spatio-temporal input. Fig. 4 shows the detailed information of the tiny version of Video Swin Transformer. Meanwhile, Doshi et al. [94] proposed a Semantic Transformer specifically designed for action recognition tasks, named Semantic Video Transformer for Action Recognition (SeViTAR). SeViTAR enhances the model's robustness by mapping the visual features extracted by video Transformers into more robust visual–semantic representations.Furthermore, Jing et al. [95] introduced a Two-Pathway Vision Transformer (TP-ViT), utilizing two parallel spatial Transformer encoders as two pathways to process input videos of different frame rates and resolutions. The high-resolution pathway contains more spatial information, while the high-frame-rate pathway captures more temporal details. The outputs from these two pathways are fused and then fed into a temporal Transformer encoder for action recognition. Additionally, the authors incorporated skeleton features into the model to further enhance performance. Subsequently, Ren et al.[96] combined SlowFast [73] with Swin Transformer [93] to propose the SF-Swin model based on the Transformer framework.This model integrates the fast and slow pathways of the SlowFast network to capture spatial structures and temporal events in videos, respectively. By employing lateral connections to fuse information from both pathways, the model's performance was further improved. To address the long-tail distribution problem, the authors also introduced Smoothed Sample Loss (SSLoss).

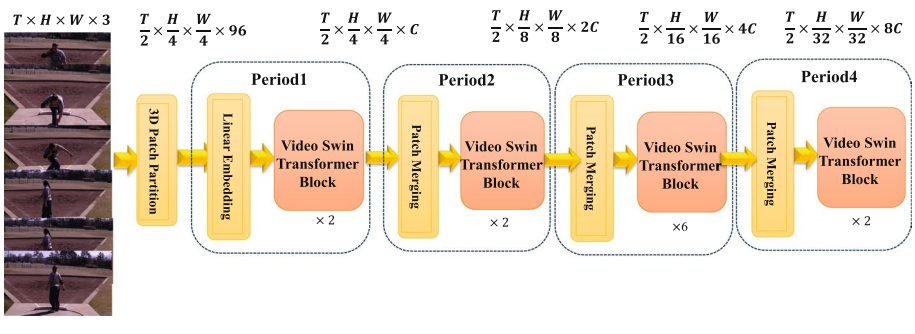

Fig. 4 The overall architecture of Video Swin Transformer

Table 1 Performance comparison on RGB-based methods for HAR

Basement	Method	Metric(%)[1]	Strength	Weakness
Hand	Dense trajectories[25]	94.2 (KTH [97])	Robust dense trajectories MBH suppresses camera motion multi-feature fusion	Limited action modeling; high computational cost; vulnerable to textureless extremes.
	Improved trajectories[26]	91.2 (UCF50 [98])	Trajectory optimization	Rely on human detector accuracy.
	Stack fisher vectors[27]	93.77 (Youtube [99])	Effective dimensionality reduction	Computational overhead
	MIFS [28]	94.4 (UCF50 [98])	Multi-scale feature capture	Limitations in complex scenarios
2DCNN	Two-Stream [30]	88.0 (UCF101 [33])	Superior performance with optical flow	Lack of trajectory-aware pooling
	Wang et al. [32]	93.2 (ActivityNet [100])	High data efficiency	Ensemble complexity
	Wang et al. [34]	91.4 (UCF101 [33])	Effective overfitting mitigation	Training strategy sensitivity
	Feichtenhofer et al. [36]	93.5 (UCF101 [33])	Effective spatial–temporal fusion	Requires optical flow preprocessing
	TS-D3D [37]	96.0 (UCF101 [33])	Dual-stream + 3D DenseNet integration	Higher computational complexity
	P-CNN [38]	79.5 (JHMDB [101])	Pose+appearance+motion integration	Requires accurate pose tracking
	TSN [39]	94.2 (UCF101 [33])	Long-range temporal modeling	Complex preprocessing for cross-modality
	Umran et al. [40]	97.1[2] (UCF101 [33])	Leveraging ROI extraction	Depends on pre-trained I3D

Table 1 (continued)

Basement	Method	Metric(%)[1]	Strength	Weakness
3DCNN	C3D [43]	90.4 (UCF101 [33])	Simple network and strong generalization	Limited long-term temporal dependency
	I3D [41]	98.0 (UCF101 [33])	Two-stream fusion	Requires optical flow preprocessing
	ResNet3D [47]	94.5 (UCF101 [33])	Depth scalability	Heavy data dependency
	STCNet [51]	96.5 (UCF101 [33])	Spatio-temporal channel modeling	Sensitivity to input configuration
	P3D [52]	93.7 (UCF101 [33])	Balanced model size and speed	Dependence on large-scale pre-training
	R(2+1)D [53]	97.3 (UCF101 [33])	Effective spatio-temporal factorization	Flow method limitation
	TrajectoryNet [54]	79.8 (Kinetics-400[102])	Explicit motion modeling	Layer-level limitations
	MiCTNet [54]	94.7 (UCF101 [33])	Cross-domain residual connections	Architectural complexity
	ARTNet [56]	94.3 (UCF101 [33])	Hierarchical multi-scale modeling	Architectural depth limitations
	S3D [57]	96.8 (UCF101 [33])	Top-heavy design	Higher computation than 2D models
	Murugan [60]	82 (UCF101 [33])	Spatio-temporal feature extraction	Lack of real-time performance metrics
	T3D [62]	93.2 (UCF101 [33])	Variable temporal modeling	Limited temporal context
	Non-local [64]	83.8 (Kinetics-400[102])	Direct long-range modeling	Sensitivity to layer placement
	Faster [66]	96.9 (UCF101 [33])	Effective aggregation	Complex model
	CapsNet [67]	98.6 (UCF101 [33])	Multi-stream fusion	Dependency on multi-stream preprocessing
	NL-MFN [68]	88.77 (UCF101 [33])	Integrating the non-local mechanism	Limited modality
	MTA-CNN [69]	89.1 (UCF101 [33])	Leveraging time-asymmetric convolutions Integrating both local/non-local features	Speed–complexity trade-off
	SlowFast [73]	81.8 (Kinetics-600[103])	Dual-pathway design Lightweight fast pathway Biological inspiration	Architectural complexity Dependency on temporal stride design

Table 1 (continued)

Basement	Method	Metric(%)[1]	Strength	Weakness
RNN	LRCN [76]	86.86 (UCF101 [33])	Temporal dependency modeling	Inconsistent class performance
	Beyond short snippets[77]	88.6 (UCF101 [33])	Long-range temporal modeling	Optical flow dependency
	Li et al. [78]	90.8 (UCF101 [33])	Integration of LSTM for Temporal cues	Static fusion weights
	Video-LSTM [79]	92.2 (UCF101 [33])	Action-based attention	Background dependency
	Lattice-LSTM [80]	93.6 (UCF101 [33])	Local superposition for non-stationary actions	Relies on pre-trained CNN features
	Three-stream LSTM[84]	99.0 (UCF101 [33])	Three-stream architecture for multi-scale spatio-temporal modeling	Increased architectural complexity
	Hierarchical LSTM[85]	96.5[3] (HAPT [104])	ST[4] feature extraction via CNN-LSTM	Limited age range in dataset affects generalization
	LSTM CrossRWKV[86]	90.83 (Jester [105])	Frame-by-frame processing	Edge detection reliance
	FAST-GRU [87]	96.9 (UCF101 [33])	Spatio-temporal resolution preservation	Long-term sequence sensitivity
Transformer	TimeSformer [88]	82.2 (Kinetics-600[103])	Convolution-free design	Heavy pre-training dependence
	VTN [89]	93.2 (Kinetics-400[102])	Strong generalization	Short-clip performance gap
	GCNet [90]	76 (Kinetics-400[102])	Effective overfitting mitigation	Training strategy sensitivity
	CFDL [36]	90.74 (PPMI [106])	Communication efficiency	Dependency on client selection
	Video Swin transformer [92]	85.9 (Kinetics-600[103])	Spatio-temporal efficiency	Computational overhead in deep layers
	SeViTAR [94]	90.54 (UCF101 [33])	Robust semantic–visual fusion	Data constraints
	TP-ViT [95]	89.38 (FineGym-288 [107])	Multi-pathway and multi-stream fusion	Modality-specific tuning
	SF-Swin [96]	85.02 (Kinetics-400 [102])	Efficient transformer integration	Hyperparameter dependency

[1] Here we select the best accuracy among the validated datasets.

[2] The result is for ROI video recognition in "Archery" class.

[3] The result is for 5 fundamental action classes.

[4] "ST" refers to spatio-temporal

2.2 Depth-based

In comparison to RGB videos, the depth modality exhibits a robust insensitivity to variations in illumination, maintains invariance to alterations in color and texture, and demonstrates reliability in estimating body silhouettes and skeletons. Additionally, it provides extensive 3D structural information of the scene. Table 2 shows the comparison of the methods for depth-based HAR algorithms.

2.2.1 Classical methods

Depth data can be interpreted as a spatio-temporal representation of human actions, capturing both spatial configurations (e.g., body pose) and temporal dynamics (e.g., motion trajectories).Yang et al.[108] proposed an architecture assembling the low-level poly-normals into the super normal vector to capture the information from depth maps. Oreifej et al.[21] proposed HON4D, describing the depth sequence with a histogram capturing the distribution of the surface normal orientation in the 4D space of time. Unlike HON4D locates on essential moving objects, Liu et al.[109] proposed SDM-BSM, depicting the silhouettes induced by the lateral component of the scene action parallel to the image plane. Rehmani et al.[110] proposed a method combing the discriminative information derived from both depth images and 3D joint positions, aiming to attain high accuracy in action recognition. Ji et al.[111] presented a method which designed a spatio-temporal cuboid to capture the geometry cues and temporal information and then combined the features extracted from the projected motion planes. And Yang et al. [112] introduced the Depth Motion Map (DMM) methodology for projecting the spatio-temporal depth structure onto action history maps. More contemporary representations of motion history maps apply a series of Histogram of Gradients (HoG) features to depict actions. Classical methods for depth-based HAR methods limited their ability to capture complex spatio-temporal patterns and generalized poorly to diverse scenarios. Traditional methods struggled with robustness to noise, viewpoint variations, and occlusions, while also underutilizing the rich 3D geometric and motion cues inherent in depth data.

2.2.2 Deep learning methods

CNN-based methods. Wang et al. [113] proposed a method combining weighted hierarchical depth motion maps (WHDMM) with three-channel deep convolutional neural networks (3ConvNets) for HAR from depth data, specifically designed for training on small datasets. They put up three tips, first rotating 3-D depth map points mimics different viewpoints, synthesizing more data and enhancing ConvNets' view-tolerance. And second WHDMMs at various temporal scales changed spatio-temporal motion patterns into 2D structures, further enhanced by pseudocolor images for recognition. Finally the three ConvNets are initialized with ImageNet models and fine-tuned on color-coded WHDMMs in three orthogonal planes. Inspired by the former work, Hou et al.[114] proposed STSDDI, a hierarchical bidirectional rank pooling method to exploit the spatio-temporal-structural information contained in the depth sequence in both spatial and temporal domain and extended the method from one person action recognition to two person recognition. Rahmani et al. [115] proposed a new method, which encompasses two distinct stages: (i) the acquisition of a generalized view-invariant human pose model through the utilization of synthetic depth images, and (ii) the modeling of temporal variations in actions. To enhance the training dataset for CNNs, they synthetically generated training data by aligning realistic synthetic 3D human models with actual motion capture (mocap) data and subsequently rendering each pose from a number of viewpoints. For the purpose of spatio-temporal representation, they employed the Group Sparse Fourier Temporal Pyramid, which

encodes the action-specific and discriminative output features of the proposed human pose model.

2.3 Skeleton-based

In contrast to RGB and depth, skeleton data cover the positional coordinates of human joints, which could be regarded as relatively advanced features for human action recognition. Skeleton sequences encode the changing trajectories of human body joints, thereby accurately characterizing informative human motions. Consequently, skeleton data are treated as a suitable modality for HAR in academic research. Currently, there are mainly four approaches for skeleton-based human action recognition, including CNN-based, RNN-based, GNN/GCN-based and Transformer-based. The comparison between the skeleton-based methods is shown in Table 3.

2.3.1 CNN-based

The methodology transforms skeleton sequences into images, where spatio-temporal dynamics are captured via color and texture features.The skeleton data could be obtained by applying pose estimation on RGB videos [116] or depth maps[117]. Du et al. [118] proposed an end-to-end hierarchical architecture for skeleton-based HAR with CNN. Initially, they represented a skeleton sequence as a matrix by concatenating the joint coordinates across each time instant and arranging these vector representations sequentially in chronological order. Subsequently, this matrix was converted into an image format and normalized to address the issue. Wang et al. [119]introduced a straightforward methodology for encoding spatio-temporal information contained within 3D skeleton sequences into a series of 2D images, termed Joint Trajectory Maps (JTMs). Furthermore, CNNs are taken to extract discriminative features for the purpose of real-time human action recognition.Wang et al.[120] proposed DST-HCN, designing a time-point hypergraph(TPH) to learn relationships in the temporal domain. By integrating a combination of multiple spatial static hypergraphs and dynamic topological propagation hierarchies (TPH), the proposed network architecture enables the model to capture comprehensive spatio-temporal representations. Inspired by observation that the categorization of an action is predicated upon the localized movements of joints, Zhu et al. [121] proposed a cuboid model for skeleton-based action recognition. Specifically, a strategy for arranging cuboids is formulated to systematically organize the pairwise displacements among all body joints, thereby obtaining a cuboid-based representation of the action, as shown in Fig. 5. Besides, Li et al. [122] proposed a straightforward yet

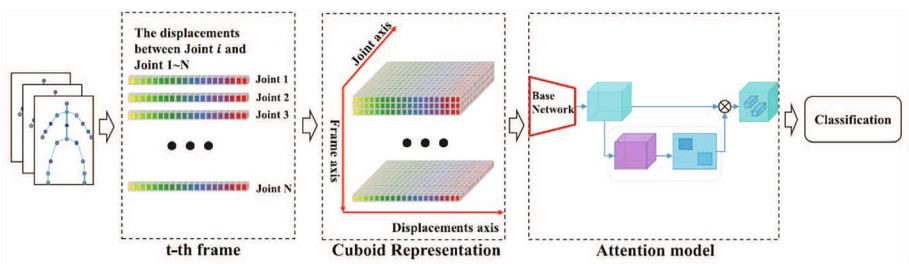

Fig. 5 The flowchart of the cuboid action representation. Figure from [121].

effective method which was introduced for encoding the spatio-temporal information of skeleton sequences into color texture images, termed Joint Distance Maps (JDMs). CNNs are then applied to extract discriminative features from these JDMs for the purpose of recognizing human actions and interactions. Ke et al. [123] proposed SkeletonNet for skeleton-based 3D action recognition. They extracted body-part-based features from each frame of the skeleton sequence and changed the features into images and fed them to the proposed network containing two parts: one for extracting general features from the input information while the other for generating a module is designed to extract general features from the input images, whereas the other is responsible for generating a discriminative and compact representation specifically for action recognition. Unlike the methods mentioned above, Rahmani et al. [115] introduced a methodology for generating such data by aligning synthetic 3D human models with authentic motion capture data and subsequently rendering the human poses from many perspectives, making it capable of generalizing effectively to real depth images depicting unseen poses, eliminating the necessity for re-training or fine-tuning procedures.

2.3.2 RNN-based

In this category of methodologies, skeletal features are taken as inputs to a RNN for the purpose of capturing and exploiting temporal dynamics. Du et al. [124] proposed an end-to-end hierarchical RNN for skeleton-based action recognition. In contrast to treating the entire skeleton as an input, they segmented the human skeleton into five distinct parts based on human physiological structure, and subsequently input these segments into five separate sub-networks. This methodology explicitly encodes spatio-temporal and structural information into a high-level representational framework. What's more, Pan et al. [125] proposed an algorithm converted the original coordinate system for raw skeleton sequences into the pose coordinate system and the trajectory coordinate system. They took a two-stream LSTM network to describe the changes of the temporal sequences. The whole architecture of the proposed two-stream RNN is shown in Fig. 6. Veeriah et al. [126] introduced a differential gating mechanism for the LSTM neural network, focusing specifically on the variations in information gain resultant from prominent motions observed between consecutive frames. Zhang et al. [127] proposed EleAtt-RNN, a simple yet effective Element-wise-Attention Gate (EleAttG) which could be easily added to an RNN block for the human action recognition task.

2.3.3 GNN/GCN-based

According to the expressive power of graph structures, it has attracted much attention for the analysis of graphs with learning models. As is known to all, skeleton data are born with the forms of graphs. Therefore, merely representing skeleton data as vector sequences processed by RNNs or as 2D/3D maps processed by CNNs is capable of fully capturing the elaborate spatio-temporal configurations and correlations among body joints. This emphasized the potential suitability of topological graph representations for modeling skeleton data.

GNN is a connectivity model that captures the inter-dependencies within a graph by helping message passing among its nodes. Si et al. [128] proposed a novel model with spatial reasoning and temporal stack learning(SR-TSL) for skeleton-based action

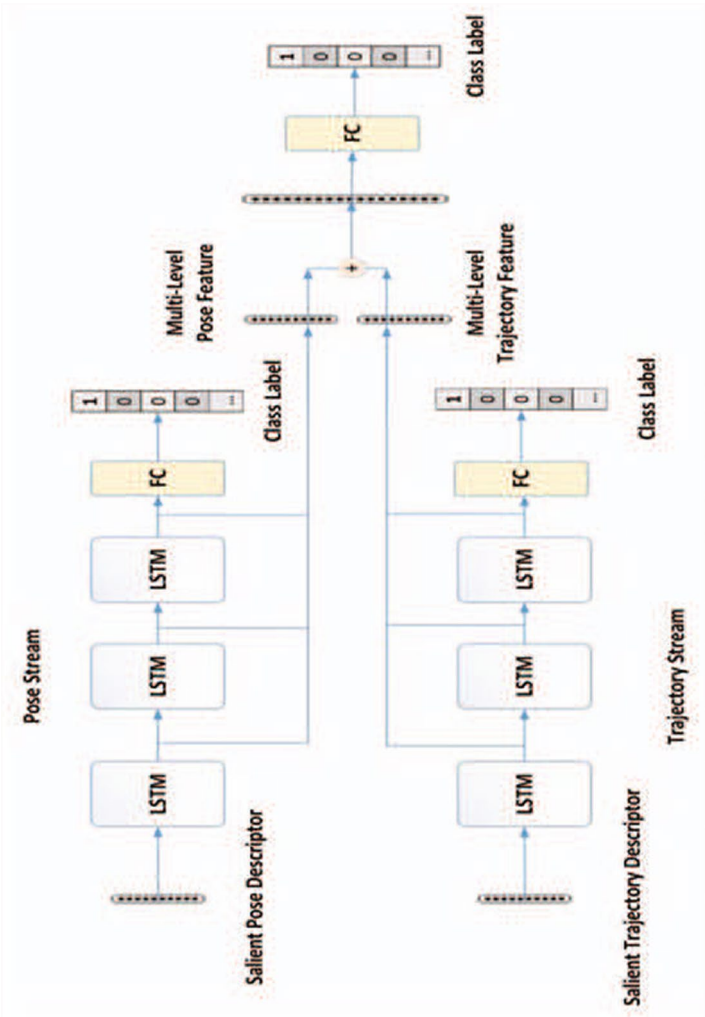

Fig. 6 The architecture of the proposed two-stream RNN. Figure from [125]

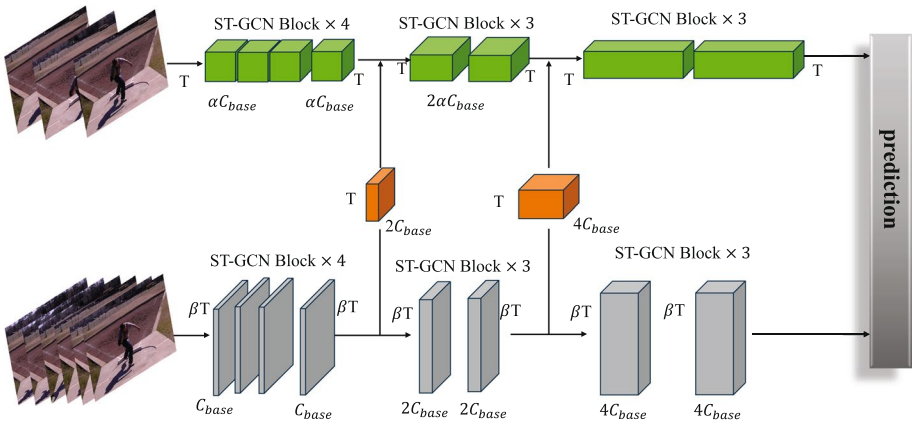

Fig. 7 The architecture of SlowFast-GCN

Table 2 Performance comparison on depth-based methods for HAR

Basement	Method	Metric(%)[1]	Strength	Weakness
Classical	SNV [108]	98.89 (MSRActionPairs3D[136])	Adaptive pyramid structure	Sensor dependency
	HON4D [21]	96.67 (MSRActionPairs3D[136])	Joint 4D spatio-temporal model	Limited generalization to non-depth data
	SDM-BSM [109]	89.5 (NHA[137])	Complementary feature fusion	Limited feature scope
	Rehmani et al.[110]	95.29 (MSRGesture3D[138])	Multi-feature fusion	Residual dimensionality
	Ji et al. [111]	98.33 (MSRActionPairs3D[136])	Multiview ST model[2]	Sensitivity to self-occlusion
	DMM-HOG [112]	98.7 (MSRAction3D[139])	Compact feature representation	Sensitivity to similar actions
Deep learning	Wang et al. [113]	90.91[3] (UTKinect-action[140])	Viewpoint and speed tolerance	Overfitting risk on small data
	STSDDI [114]	96.58 (G3D[141])	Hierarchical feature modeling	Complex object interaction limitations
	Rahmani et al.[115]	92.0 (N-UCLA[142])	View-invariant representation	Segmentation dependency

[1] Here we select the best accuracy among the validated datasets.

[2] "ST" refers to spatio-temporal.

[3] The result is for the second highest score

recognition, consisting of a spatial reasoning network(SRN) and a temporal stack learning (TSLN). The SRN employs a residual graph neural network (RGNN) to capture high-level spatial structural information within individual frames, whereas the TSLN took a composition of multiple skip-clip LSTMs to model the elaborate temporal dynamics of skeleton sequences. Besides, Shi et al. [129] introduced a brand-new directed GNN to obtain the information of joints, bones and the relationship between them. Then it could make precise prediction with the information acquired before.

For GCN-based methods, there are a huge number of algorithms for skeleton-based human action recognition. First of all, Yan et al. [130] introduced a novel model for changing skeletons, named Spatial–Temporal Graph Convolutional Networks

(ST-GCN). This model surpassed the constraints of prior methodologies by autonomously acquiring both spatial and temporal patterns directly from the data. This formulation not only enhances expressive power but also bolsters generalization capability, thereby conferring superior performance in various contexts. Ye et al. [131] proposed a novel joints relation-reasoning, graph convolutional network(JRR-GCN). In contrast to conventional methods based on spatial convolutional networks, the adjacency matrices in JRR-GCN are automatically inferred by the Joints Relation-Reasoning Network (JRR). This approach leads to the generation of a more realistic representation of skeleton topology and produces superior adjacency matrices for each individual sample. Besides, Lin et al.[132] presented a brand-new architecture termed SlowFast-GCN, which integrates the strengths of ST-GCN and SlowFastNet with the incorporation of active human skeletons. This innovative approach aims to enhance the precision of human action recognition, and the overall architecture of this network is shown in Fig. 7. What's more, Yang et al. [133] proposed a hybrid network architecture, termed HybridNet, aiming to smoothly integrate GCNs with CNNs. The HybridNet effectively absorbed structural information while accurately modeling the elaborate relationships among inter-frame joints. Zang et al. [134] introduced SparseShift-GCN, which first applied the Conv-Shift-Conv(CSC) module and the Shift-Conv-Shift-Conv(SC2) module to take the place of the Shift-Conv-Shift(SCS) module in spatial graph convolution of Shift-GCN, respectively. And it proposed the substitution of the shift module in the original Shift-GCN with a sparse shift module, terming the resultant architecture SparseShift-GCN. To better capture the long-range dependency, Qiu et al. [135] proposed a novel unsupervised method called Global–Local Temporal Attention Graph Convolutional Network(GLTA-GCN), comprising two distinct branches: a local branch and a global branch. Both branches employ graph convolution units alongside a self-attention mechanism to enhance the extraction of spatio-temporal features. What's more, two specifically designed loss functions are incorporated to constrain the model, ensuring it extracts more critical local joint features while preserving intrinsic structural information.

2.3.4 Transformer-based

We have introduced the skeleton-based HAR algorithms in the field of CNN-based, RNN-based and GNN/GCN-based, this section we would change the view to transformers [65]. Self-attention-based architectures are commonly employed for feature integration, owing to their proven effectiveness in capturing long-term dependencies. Kong et al. [143] proposed a multi-scale temporal transformer(MTT) for skeleton-based human action recognition, as shown in Fig. 8. Firstly, the raw skeleton data undergo embedding through graph convolutional network (GCN) blocks and multi-scale temporal embedding modules (MT-EMs), which are structured as multiple branches to promote the extraction of features across various temporal scales. Secondly, transformer encoders (TE) are incorporated to integrate these embeddings and model the long-term temporal patterns. Besides, they proposed a task-oriented lateral connection (LaC) with the objective of aligning semantical hierarchies. What's more, Zhang et al. [144] regarded transformer as a kind of active GCN, and the weight assigned to each node is dynamically ascertained based on the data. In this research, they proposed a three-dimensional

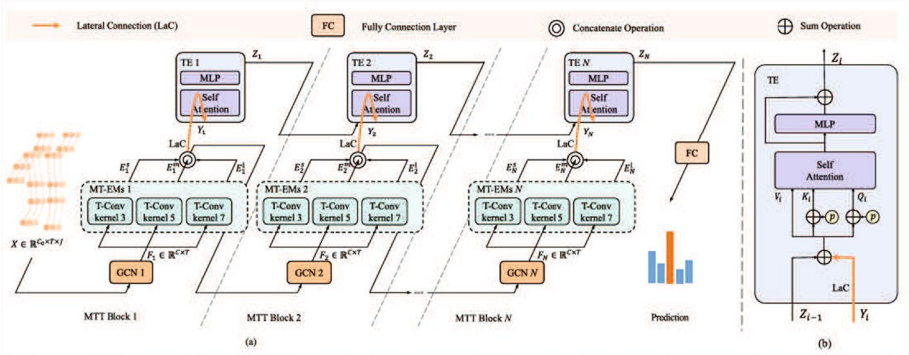

Fig. 8 The overall architecture of MTT [143]. Figure from [143]

position encoding methodology to address the challenge of representing node spatial information, thereby enabling the application of the transformer architecture to graph data. Furthermore, analogous to ST-GCN [130], they introduced the Space–Time Transformer (ST-TR), which applied transformers in both spatial and temporal domains to extract spatio-temporal features from skeleton data, ultimately promoting action recognition tasks. Jing et al. [95] presented a novel structure called Two-Pathway Vision Transformer (TP-ViT) for skeleton-based human action recognition. In this work, two parallel spatial transformer encoders were taken to constitute two distinct pathways, each processing the input video with differing framerates and resolutions. Specifically, the pathway designed for high-resolution captured a richer array of spatial information, whereas the pathway suited for high-framerate captured a more abundant temporal information.

2.4 RGB-D-based

As delineated in the preceding sections, RGB, depth, and skeleton modalities exhibit distinct and specific characteristics. Consequently, it is essential to explore effective methodologies for integrating the respective strengths of these modalities within the framework of deep learning approaches. To overcome this tough problem, a number of algorithms have been presented. There are three kinds of approaches in three subfields: CNN-based, RNN-based and Transformer-based. The detailed comparison for the RGB-D methods can be seen in Table 5.

2.4.1 CNN-based

Zhu et al.[149] integrated RGB and depth information within a pyramidal 3D convolutional network which was grounded on the C3D architecture proposed by Tran et al. [43] for the purpose of gesture recognition. They devised a methodology involving pyramid input and pyramid fusion for each modality, eventually employing late score fusion for the final recognition process. Xu et al. [150] proposed a new action tube extractor for action recognition based on RGB-D data. The methodology comprises two integral components: spatial tube extraction and temporal sampling. The initial component is grounded in MobileNet-SSD, serving the purpose of delineating the spatial confines within which the action occurs. The subsequent component leverages the Structural Similarity Index (SSIM) and is specifically designed to eliminate frames devoid of

Table 3 Performance comparison on skeleton-based methods for HAR

Basement	Method	Metric(%)[1]	Strength	Weakness
CNN	Shift-GCN [118]	96.5(CV) (NTU RGB+D[145])	Adaptive receptive fields	View variation sensitivity
	DST-HCN [120]	96.8(CV) (NTU RGB+D[145])	High-order feature fusion	Hyperparameter sensitivity
	Zhu et al. [121]	96.10 (UTKinect [146])	Robustness to multi-person interactions	Skeleton dependency
	JDM [122]	88.10 (UTD-MHAD[147])	View-invariant representation	Sensitivity to mirrored actions
	SkeletonNet [123]	93.47 (SBU [148])	Generalization to multi-person actions	Skeleton dependency
	Rahmani et al.[115]	92.0 (N-UCLA [142])	Synthetic data efficiency	Limited single-view detail
	Wang et al. [119]	96.02 (G3D [141])	Cross-view generalization	Action similarity limitations
RNN	HBRNN-L [124]	94.64 (MSRAction3D[139])	Hierarchical spatio-temporal design	Low computational efficiency
	TMFF [125]	95.96 (UTKinect [146])	Dual coordinate modeling	Hyperparameter sensitivity
	DRNN [126]	93.96 (KTH [97])	Cross-domain generalization	Long-term dependency limits
	EleAtt-RNN [127]	90.7 (N-UCLA [142])	Element-wise attention	Input noise sensitivity
GNN/GCN	SR-TSL [128]	92.4(CV) (NTU RGB+D[145])	Efficient long-term temporal modeling	Heavy reliance on data preprocessing
	DGNN [129]	96.1(CV) (NTU RGB+D[145])	Directed spatio-temporal graph modeling	Adaptive graph training constraints
	ST-GCN [130]	88.3(CV) (NTU RGB+D[145])	Joint spatial–temporal modeling	Higher computational complexity
	GRR-GCN [131]	91.20(CV) (NTU RGB+D[145])	RL for relation reasoning	Sensitivity to RL hyperparameters
	SlowFast-GCN[132]	90.0(CV) (NTU RGB+D[145])	ST-GCN backbone for graph structure	Hyperparameter sensitivity in temporal fusion
	HybridNet [133]	96.90(CV) (NTU RGB+D[145])	Multi-stream strategy and split-attention	Skeleton representation rigidity
	SparseShift-GCN[134]	96.6(CV) (NTU RGB+D[145])	Sparse shift for redundancy reduction	Loss function combination limitations
	GLTA-GCN [135]	81.2(CV) (NTU RGB+D[145])	Dual-branch global–local feature modeling TAUs[2] for long-range dependencies	Reconstruction task limitations
Transformer	MTT [143]	96.7(CV) (NTU RGB+D[145])	Transformer-driven long-term dependency	Semantic gap in early layers
	ST-TR [144]	91.8(CV) (NTU RGB+D[145])	Global ST[3] modeling with Transformers	Structural rigidity in position encoding

[1] Here we select the best accuracy among the validated datasets.

[2] TAU refers to temporal attention units.

[3] "ST" refers to spatio-temporal

significant motion from the primary action tube. Duan et al. [151] presented a spatial–temporal architecture employing a consensus-voting mechanism to explicitly capture the long-term structure of video sequences and to mitigate estimation variance in the presence of extensive inter-class variations. To mitigate distractions arising from the background, a parallel 3D depth-saliency convolutional neural network stream (3DDSN)

was integrated to discern subtle motion features. Subsequently, a score fusion approach was employed for the final recognition process. The aforementioned methods treated RGB and depth as distinct channels, with fusion occurring subsequently. Besides, Srihari et al. [152] proposed a 4-stream CNN network, which comprised two spatial RGB-D video data streams and two additional streams for apparent motion. The motion streams derived their inputs from the optical flow extracted from RGB-D videos. Each of the four CNN streams is structured with 8 convolutional layers, followed by 2 dense layers and a SoftMax layer. Besides, a score fusion model is employed to integrate the scores obtained from these four streams. What 's more, Wang et al. [153] introduced a novel deep neural network called c-ConvNet on both RGB visual features and depth visual features, which deeply assembled the two modalities for human action recognition. In contrast to the traditional ConvNet, learning deep separable features for classification based on homogeneous modalities using a single soft-max loss function, the c-ConvNet significantly enhanced the discriminative capability of the deeply learned features and mitigated the unwanted modality discrepancy. Zhou et al. [154] proposed to decouple and recouple spatio-temporal representation for RGB-D based action recognition. They divided the challenging task of learning spatio-temporal representations into three elaborate sub-tasks. Firstly, they focused on acquiring high-quality features that are independent of their dimensions, through which decoupling spatial and temporal modeling. Then, reconnecting these de-coupled representations to foster a stronger space–time dependency is a must. Lastly, a mechanism called Cross-modal Adaptive Posterior Fusion (CAPF), designing to capture comprehensive cross-modal spatio-temporal information from RGB-D data.

2.4.2 RNN/LSTM-based

Recent years, temporal convolutions and RNN have been combined to capture the temporal information. Shi et al. [155] proposed a new network trying to learn an RNN driven by privileged information (PI) in three steps. An encoder is initially pre-trained to acquire a joint embedding that integrated depth appearance with Pose Information (PI), specifically skeleton joints. Subsequently, the acquired embedding layers underwent further tuning during the learning phase, with the objective of optimizing the network by using PI in the form of a multi-task loss function. However, utilizing PI as an auxiliary task demonstrated limited efficacy in enhancing the performance of the primary task, namely classification, owing to the inherent discrepancy between the two. To address this issue, a bridging matrix was introduced in the refining stage, which serves to interconnect the two tasks by uncovering latent PI. Besides, Mahasseni et al. [156] proposed a regularization method for LSTM learning, which utilized the output of an additional encoder LSTM (eLSTM) trained on 3D human-skeleton data as the regularization term. This regularization approach is grounded on the hypothesis that, given the commonality of human motion in both videos and skeleton sequences, their respective feature representations ought to exhibit similarities. The adoption of skeleton sequences, which were independent of viewing angles and free from background noise, is anticipated to enhance the capture of crucial motion patterns associated with human body joints in 3D space.

2.4.3 Other architecture-based

In this subsection, we would provide the algorithms using other architectures based on RGB-D data for readers in order to offer a more comprehensive review of the RGB-D based mythologies. Xiao et al. [157] introduced a Shift Swin Transformer Multimodal Networks method. The design of the shift mechanism was derived from the Temporal Shift Module (TSM), enabling the exchange of information between neighboring frames through the temporal shifting of partial channels. In addition, taking the advantages of the Swin Transformer network and Temporal Segment Networks (TSN), a novel feature learning approach was introduced to enhance overall performance. The overall architecture of the model is shown in Fig. 10. What's more, Wu et al. [158] described a brand-new method named Deep Dynamic Neural Networks(DDNN) for multimodal gesture recognition. This approach learned high-level spatio-temporal representations using deep neural networks suited to the input modality. A Gaussian–Bernoulli Deep Belief Network (DBN) is employed for addressing skeletal dynamics, whereas a 3D Convolutional Neural Network (3DCNN) is taken for the management and fusion of batches comprising depth and RGB images (Fig. 9).

3 Benchmark datasets

In the past decade, numerous benchmark datasets for RGB-D information have been compiled and publicly released for the benefits of the research community. As for the RGB-D datasets, we could divided them into two categories, segmented datasets and continuous datasets. And Sect. 3 would thoroughly introduce the former two and the acquisition of RGB-D data. Table 4 summarizes the state-of-the-art algorithms and the experimental results for popular HAR datasets.

3.1 Data acquisition

RGB-D data strands for red, green, blue and depth data monitored by RGB-D sensors. An RGB-D image furnishes depth information on a per-pixel basis, which is thoroughly aligned with the corresponding pixels in the image. Specifically, the depth information

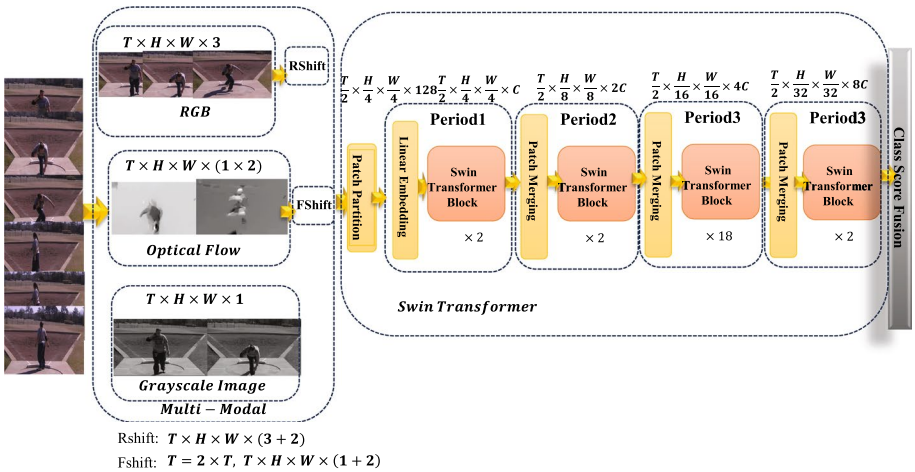

Fig. 9 The whole architecture of Shift Swin Transformer multimodal networks

Table 4 Details of the public datasets for HAR

Dataset	Year	Seg/Con[1]	Modality[2]	*Classes[3]	*Subjects	*Samples	SOTA[4]	Metric(%)[5]
CMU Mocap[159]	2001	Seg	RGB,S	45	144	2,235	ODMO[160]	88.3
KTH[97]	2004	Seg	RGB	6	25	2391	CNN - GRU[161]	95.38
MSR-Action3D[139]	2010	Seg	S,D	20	10	567	DAM[162]	94
HMDB51[163]	2011	Seg	RGB	51	-	6,766	Video MAEV2[164]	88.7
UCF101[33]	2012	Seg	RGB	101	-	13,320	FTP-UniformerV2 [165]	99.7
Sports-1M[166]	2014	Seg	RGB	487	-	1,113,158	ip-CSN-152[167]	75.5[6]
ChaLearn2014[168]	2014	Con	RGB,D,S,Au	20	27	13,858	CTC[169]	98.2
N-UCLA[142]	2014	Seg	RGB,D,S	10	10	1475	DVANet[170]	94.4(CS) 96.5(CV)
ActivityNet[100]	2015	Seg	RGB	203	-	27,801	Text4Vis[171]	96.9
Charades[172]	2016	Con	RGB	157	267	9848	MSQNet[173]	47.57[7]
NTU RGB+D[145]	2016	Seg	RGB,D,S,IR	60	40	56,880	DSCNet[174]	97.4(CS) 99.4(CV)
PKU-MMD[175]	2017	Con	RGB,D,S,IR	51	66	1076	DSCNet[174]	97.4(CS) 98.8(CV)
THU-READ[176]	2017	Seg	RGB,D	40	8	3840	UMDR[177]	90.04
Kinetics-400[102]	2017	Seg	RGB	400	-	306,245	Florence[178]	86.5
Something-Something-thingv1[179]	2017	Seg	RGB	174	-	108,499	InternVideo[180]	70
NTU RGB+D 120[181]	2019	Seg	RGB,S,D,IR	120	106	114,480	DSCNet[174]	96.7(C-setup) 95.6(C-sub)
ETRI-Activity3D[182]	2020	Seg	RGB,S,D	55	100	112,620	FSA-CNN[182]	93.7
UAV–Human[183]	2021	Seg	RGB,S,D,IR	155	119	67,428	PMI Sampler[184]	55
	2022	Seg	RGB,S,D,IR,#,C	12	41	217	EGT	93.51±.04
	2023	Seg	RGB	21	-	50000	UniPose	92.99±.28
WiFall[185]	2024	Seg	WiFi	5	10	1000	CSI-BERT2[186]	88.59

[1] Seg: segmented, Con: continuous

[2] D: depth, S: skeleton, Au: audio, Ac: accelerometer, IR: IR videos

[3] *: numbers

[4] SOTA: state-of-the-art algorithm right now

[5] Without specific notation, accuracy is used for Top-1 accuracy. And here we select the best accuracy among the modalities contained in the dataset.

[6] Top-5 accuracy is 92.8%.

[7] Here refers to mAP

makes up an image channel where each pixel represents the distance from the image plane to the corresponding object depicted in the RGB image.

3.1.1 RGB-D data acquisition

The acquisition of depth information primarily relies on triangulation and time-of-flight (ToF) techniques. Triangulation could be implemented passively through stereovision, which involves capturing the same scene from multiple perspectives to derive depth information. This method mimics the human vision principle, where depth perception is computed based on the disparity between images captured from differing perspectives. However, this approach necessitates a thorough understanding of camera geometry and requires calibration whenever there is a change in system configuration. Alternatively, an active method exploits structured light, which projects an infrared (IR) light pattern onto the scene to estimate depth by analyzing the variations in the pattern caused by differing object depths. What's more, ToF and Light Detection and Ranging (LiDAR)

Fig. 10 The view of Kinect v1, Kinect v2, Intel RealSense Camera and OrbbecAstra Pro

Table 5 Performance comparison on RGB-D-based methods for HAR

Basement	Method	Metric(%)[1]	Strength	Weakness
CNN	Pyramidal 3DCNN[149]	50.93 (ChaLearn IsoGD[191])	Multi-scale pyramid input	Limited adaptability to dynamic hand scales and viewpoints
	ATE 3DCNN [150]	93.56(CS) (NTU RGB+D[145])	Action tube extractor (ATE) Two-stream I3D architecture	Vulnerable to occlusions
	Duan et al. [151]	97.83 (RGBD-HuDaAct[192])	3D Depth-Saliency Network (3DDSN)	Fine-grained limitations
	c-ConvNet [152]	96.86 (BVCAction3D[193])	Cooperative training paradigm	Elevated computational complexity
	Wang et al. [153]	92.0 (N-UCLA [142])	Synthetic data efficiency	Limited single-view detail
	Zhou et al. [154]	97.3(CV) (NTU RGB+D[145])	ST^2 decouple–recouple learning	Elevated computational complexity
RNN/LSTM	PRNN [155]	94.9 (MSRAction3D[139])	Exploitation of privileged information Three-stage training framework	Sensitivity to skeleton annotation
	RLSTM [156]	86.9 (UCF101 [33])	Multimodal skeleton regularization	Skeleton-class overlap limitations
Others	SST^2 [157]	97.4 (UCF101 [33])	ST^3 Modeling via shift operations	Grayscale image information loss
	DDNN [158]	86.4 (ChaLearn LAP[194])	Multimodal complementary fusion	Preprocessing dependency

[1] Here we select the best accuracy among the validated datasets.

[2] "SST" stands for Shift Swin Transformer.

[3] "ST" refers to spatio-temporal

scanners both measure the duration of time it took for light to strike an object's surface and return to the detector. While LiDAR employs mechanical components to scan its surroundings, ToF utilizes integrated circuits to perform distance computations.

3.1.2 RGB-D sensors

In the field of consumer RGB-D sensors, the majority rely on either structured light or time-of-flight (ToF) methodologies. These RGB-D sensors inherently possess noise and data distortions, which are addressed through the application of specially designed algorithms. Notably, ToF sensors offer superior depth resolution compared to others, achieving accuracy within a few millimeters. Conversely, structured light systems are less effective in outdoor environments due to the remarkable interference caused by solar light on infrared (IR) cameras. For HAR tasks that do not need extremely high depth resolution and precision, both structured light sensors and ToF devices have been successfully implemented. These devices offer a favorable balance between cost, performance, and usability, enabling the development of unobtrusive and privacy-preserving solutions. In the following subsections, we outline some of the consumer-preferred RGB-D sensors.

Intel® RealSense™ depth cameras. The Intel Real-Sense depth camera series makes up a comprehensive family of stereoscopic and portable RGB-D sensors, characterized by their subpixel disparity accuracy, integrated assisted illumination, and generalized performance, even in outdoor environments. Keselman et al. [187] presented a thorough overview of the stereoscopic Intel RealSense RGB-D imaging systems. The R400 family acts as the successor to the R200 family, incorporating advancements in its stereoscopic matching algorithm and correlation cost function. In addition, design optimizations have been implemented, enabling the R400 family to operate at lower power consumption compared to the R200 family while maintaining the same image resolutions, as shown in Fig. 10.

Microsoft® RealSense™ sensors. As an advanced company in artificial intelligence, Microsoft has released the Kinect RGB-D sensors, a low-cost but high-resolution tool. And the signals from the sensors are capable of manipulating by common academic practices. There are two kinds of Kinect sensors, Kinect sensor V1 and Kinect sensor V2, the former one uses structured light and the latter one is supplemented by ToF. From [188], we could know that the former one contains two CMOS sensors, one for RGB imaging and another for depth sensing, both of which are in combination with a laser projector. And the horizontal field of view measures 57 degrees, while the vertical field of view spans 43 degrees [189]. When it comes to the Kinetic v2, it exhibits superior overall precision, responsiveness, and intuitive functionalities, promoting the accelerated development of applications capable of responding to movements, gestures, and voice commands. What's more, the v2 sensor's color camera has been enhanced to provide full 1080p video resolution, which can be displayed with a one-to-one correspondence to the resolution of the viewing screen[188]. Figure 10 shows the details of Kinect v1, Kinect v2, Intel RealSense Camera and OrbbecAstra Pro.

Orbbec® Depth cameras. Orbbec Astra sensors incorporate a processor that eliminates the necessity for a traditional cable-based connection to the sensor for data transmission. The Orbbec sensor, analogous to the Kinect, comprised a RGB camera, a depth camera, an infrared (IR) projector, and two microphones [190]. It could be used with the Astra SDK, or the OpenNI framework, or other third-party SDK. The device features a depth camera with a resolution of 640x480 pixels and an RGB camera with a resolution of 1280x960 pixels, both capable of capturing images at a rate of 30 frames

per second. It utilizes USB 2.0 as the data interface for transmitting this imagery. We could see the details of Orbbec Depth Camera in Fig. 10.

3.2 Segmented datasets

Segmented datasets refer especially for those in which individual samples represent complete actions or gestures, from their inception to completion, with each segment corresponding to a distinct action. These datasets are primarily utilized for classification tasks. Below, we enumerate several segmented datasets that are commonly for the evaluation of deep learning-based methodologies. Table6 and Table 6(Continued) show the performance of the SOTA methods on some widely used segmented datasets in detail, the bold methods are the state-of-the-art method for certain dataset.

3.2.1 CMU Mocap

CMU Graphics Lab Motion Capture Database(CMU Mocap)[159] stands as one of the most pioneering sources of comprehensive data which cover a diverse range of human actions. This includes interactions between dual subjects, human locomotion, engagements with uneven terrain, various sports, and other forms of human behavior. This dataset comprises both RGB and skeleton modalities.However, its utility requires careful consideration of technical limitations: (1) data quality is constrained by occlusion-induced skeletal data gaps in complex movements and RGB videos captured in controlled lab settings (640×480 resolution, limited lighting and camera angles), which may hinder real-world generalization; (2) annotation accuracy relies on marker-based skeletal tracking with potential millimeter-scale errors during rapid motions, while action labels are manually annotated by experts without reported inter-annotator agreements, contrasting with standardized protocols in datasets; and (3)applicability is scenario-dependent—its strength in multi-agent interactions suits social behavior studies, but lower-resolution RGB data limit tasks requiring fine-grained visual analysis, where datasets like NTU RGB+D [145] offer higher-resolution 2D/3D annotations. Thus, while CMU Mocap remains valuable for understanding naturalistic human dynamics, its deployment should be guided by its specific trade-offs between diversity, precision, and modality richness.

3.2.2 KTH

The KTH dataset[97] emerges as a preeminent benchmark within the domain of action recognition, including six definitive actions: walking, jogging, running, boxing, hand-waving, and hand-clapping. To precisely capture the subtle distinctions of performance, each action is executed by a varied cohort of 25 individuals, with systematic variations thoroughly introduced for each action per performer. These variations span outdoor settings (designated as s1), outdoor settings with scale variations (s2), outdoor settings exhibiting diverse attire (s3), and indoor environments (s4). These redundant scenarios pose a formidable challenge to the algorithmic capacity to discern actions, irrespective of the background clutter, the appearance of the actors, and their scale, thereby furnishing a stringent test of algorithmic robustness and adaptability. However, its utility requires critical evaluation of three aspects: first, the data quality is constrained by a monocular camera setup (640×480 resolution, 25 fps), which lacks depth perception

Table 6 SOTA performance comparison on some widely used segmented datasets

Dataset	Method	Year	Metric(%)	Core features
KTH[97]	KMP[195]	2012	90.2	Early 3D CNNs had high complexity and low accuracy due to heavy computation.
	CNN-LSTM[196]	2020	93.86	Combining 2D CNN for spatial features and LSTM for temporal modeling.
	Differential RNN[197]	2015	93.96	Uses differential operations in RNN to model temporal differences.
	3D-ConvNet+LSTM[198]	2011	94.39	3D conv-based early spatio-temporal model: high computational cost.
	SIFT+OF+CNN[193]	2020	94.96	Fuses SIFT and optical flow with CNN, relying on multi-modal feature integration.
	CNN-GRU [161]	2024	95.38	Training-free: unsupervised frame selection + lightweight GRU, efficient and accurate.
HMDB51[163]	R2+1D-BERT[199]	2020	85.10	BERT+ R(2+1)D
	SO+MaxExp+IDT[200]	2024	85.70	High-order tensor pooling with HDP/EPN
	SCK[201]	2022	86.11	Two tensor-based feature representations
	TO+MaxExp+IDT[200]	2024	87.21	High-order tensor pooling with HDP/EPN
	DEEP-HAL with ODF+SDF[202]	2021	87.56	Taking RGB frames as input to learn to predict both action concepts and auxiliary descriptors.
	VideoMAE V2-g[164]	2023	88.7	VideoMAE enables efficient billion-parameter video foundation models via dual masking and progressive training.
UCF101[33]	OmniSource[203]	2020	98.6	OmniSource unifies image-video formats for webly-supervised learning.
	SMART[204]	2020	98.64	Selecting good frames helps in action recognition performance even in the trimmed videos domain.
	BIKE[205]	2023	98.8	Using cross-modal bridges for bidirectional video-text knowledge to enhance video representation.
	OmniVec2[206]	2024	99.6	Modality-specialized tokenizers, shared transformer with cross-attention, and task heads enable unified multimodal multitask learning.
	VideoMAE V2-g[164]	2023	99.6	VideoMAE enables efficient billion-parameter video foundation models via dual masking and progressive training.
	FTP-UniFormerV2-L/14 [165]	2024	99.7	FTP combines ViTs and VLMs with four processors for enhanced video action representation.

Table 6 (continued)

Dataset	Method	Year	Metric(%)	Core features
Sports-1M[166]	R(2+1)D-Flow-32frame[53]	2018	68.4(Top-1) 88.7(Top-5)	2D Conv for spatial and 1D Conv for temporal with flow modality.
	Conv pooling [207]	2015	71.7(Top-1) 90.4(Top-5)	Two methods for long-term video analysis: convolutional temporal pooling and LSTM-based sequence modeling.
	R(2+1)D-RGB-32frame[53]	2018	73(Top-1) 91.5(Top-5)	2D Conv for spatial and 1D Conv for temporal with RGB modality.
	R(2+1)D-Two-stream-32frame[53]	2018	73.3(Top-1) 91.9(Top-5)	2D Conv for spatial and 1D Conv for temporal with two streams.
	ip-CSN-101 [167]	2019	74.9(Top-1) 92.6(Top-5)	Study 3D group convolutions for video classification, showing channel separation boosts efficiency and accuracy via CSN with ResNet101.
	ip-CSN-152 [167]	2019	75.5(Top-1) 92.8(Top-5)	Study 3D group convolutions for video classification, showing channel separation boosts efficiency and accuracy via CSN with ResNet152.
N-UCLA[142]	DA-Net[208]	2018	92.1(CS) 86.5(CV)	DA-Net combines shared/view-specific representations and CRF-based fusion for multi-view action recognition.
	Glimpse Clouds[209]	2018	-(CS) 87.6(CV)	Pose-free RGB action recognition via attention-driven glimpse sequence prediction.
	RL-NET [210]	2020	87.5(CS) 83.1(CV)	Unsupervised RGB-based multi-view action recognition via unseen viewpoint prediction and scene dynamics encoding.
	ViewCLR [211]	2023	-(CS) 89.1(CV)	ViewCLR learns self-supervised video representation invariant to camera viewpoint changes.
	ViewCon [212]	2023	-(CS) 91.7(CV)	Supervised contrastive framework with synchronized viewpoints and classifier-guided hard negatives enhances robust multi-view action recognition.
	DVANet[170]	2024	94.4(CS) 96.5(CV)	Transformer-based disentanglement with contrastive losses for robust multi-view action recognition

Table 6 (continued)

Dataset	Method	Year	Metric(%)	Core features
ActivityNet[100]	DSANet [213]	2021	90.5	DSA module enables adaptive long-range temporal aggregation for clip-based models.
	TSQNet [214]	2022	93.7	TSQNet introduces class-aware temporal saliency queries with cross-modal fusion for efficient video recognition.
	NSNet [215]	2022	94.3	NSNet suppresses non-salient frames via dual supervision and multi-granularity saliency fusion.
	InternVideo2-6B[216]	2024	95.9	InternVideo2: 6B-parameter progressive-training video foundation model with cross-modal alignment.
	BIKE [205]	2023	96.1	BIKE uses cross-modal bridges for bidirectional video-text knowledge to enhance video representation.
	Text4Vis[171]	2022	96.9	Text4Vis: Leverages pre-trained language models for knowledge transfer in video classification
NTU RGB+D[102]	MMNet [217]	2023	96.0(CS) 98.8(CV)	MMNet: model-based multimodal fusion for RGB-D action recognition using ST-GCN attention transfer.
	EPAM-Net [218]	2024	96.1(CS) 99.0(CV)	Efficient pose-driven attention-guided multimodal network with 6.2−9.9x FLOPs and 9−9.6x parameter reductions.
	UMDR [177]	2023	96.2(CS) 98.0(CV)	UMDR framework with ShuffleMix and CFCer achieves great performance for RGB-D action recognition.
	π-ViT [219]	2024	96.3(CS) 99.0(CV)	Pose-induced video transformer with 2D/3D skeleton modules for ADL action recognition.
	Pose C3D [220]	2022	97.0(CS) 99.6(CV)	PoseConv3D: 3D heatmap-based skeleton action recognition with spatio-temporal features, robustness to noise, and cross-dataset generalization.
	DSCNet[174]	2023	97.4(CS) 99.4(CV)	DSCNet: RGB-skeleton complementary action recognition via dense-sparse sampling, background suppression with STMEM, and sparse multi-scale skeletal modeling.

Table 6 (continued)

Dataset	Method	Year	Metric(%)	Core features
S-SV1[1][179]	ATM [221]	2023	65.6	ATM: arithmetic temporal module with four operations for low-cost, plug-and-play temporal modeling in CNNs/ViTs.
	Slide4Video [222]	2023	67.3	A spatial–temporal side network for 75% memory-efficient video ViT transfer, enabling 4.4B ViT-E (14x larger than ViT-L).
	InternVideo[180]	2022	70.0	General video foundation model combining masked video modeling and video-language contrastive learning for enhanced applications.
S-SV2[1][179]	Hiera-L [223]	2023	76.5	Self-Supervised Learning + Transformer + No Extra Data.
	VideoMAE V2-g[164]	2023	77.0	VideoMAE enables efficient billion-parameter video foundation models via dual masking and progressive training.
	MVD [224]	2023	77.3	MVD: Two-stage masked video distillation framework with spatial–temporal co-teaching using image/video teachers.
NTU RGB+D 120[181]	DSTSA-GCN [225]	2025	90.97(C-Setup) 88.7(C-Sub)	Group Channel-wise Graph Convolution (GC-GC) Group Temporal-wise Graph Convolution (GT-GC) Multi-Scale Temporal Convolution (MS-TCN).
	JPFormer [226]	2024	91.4(C-Setup) 89.4(C-Sub)	Joint-Partition Group Attention for skeleton action recognition, capturing multi-granularity joint-part correlations via adaptive reparameterized partitioning.
	3DA [227]	2023	91.4(C-Setup) 90.5(C-Sub)	A new 3D deformable transformer for action recognition with adaptive spatial–temporal receptive fields and a cross-modal learning scheme.
	IPP-Net [228]	2023	91.7(C-Setup) 90.0(C-Sub)	Leveraging both skeletons and human parsing feature maps in dual-branch approach.
	STAR-Transformer[229]	2023	92.7(C-Setup) 90.3(C-Sub)	RGB+Pose+ViT.
	DSCNet [174]	2023	96.7(C-Setup) 95.6(C-Sub)	DSCNet: RGB-skeleton complementary action recognition via dense-sparse sampling, background suppression with STMEM, and sparse multi-scale skeletal modeling.

[1] S-S refers to something-something.

and struggles with occlusion handling; second, action labels are manually defined by researchers without documented inter-annotator agreements, contrasting with datasets like HMDB51 [163] that employ crowdsourced annotations; finally, while the controlled

variations (e.g., lighting, actor appearance) make it suitable for evaluating environmental robustness, its minimal background clutter and absence of complex interactions (e.g., multi-person actions) [added] limit real-world generalizability compared to large-scale benchmarks like UCF101 [33] or Kinetics [102].

3.2.3 MSR-Action3D

MSR-Action3D[139] dataset, uses the advanced Kinect™ sensor technology through a collaborative effort between Microsoft Research, Redmond, and the University of Wollongong. This comprehensive dataset covers an array of 20 distinct actions, specifically: high arm wave, horizontal arm wave, hammering, hand catch, forward punch, high throw, drawing an 'X', drawing a tick, drawing a circle, hand clap, two-handed wave, side-boxing, bending, forward kick, side kick, jogging, tennis serve, golf swing, pickup, and throw. A diverse group of ten subjects executed each action a total of three times, ensuring generic data representation.All recordings were captured at 640× 480 resolution with 30 fps using a single Kinect sensor, with post-processing applying frame-difference algorithms to remove static backgrounds. All recordings were captured from a fixed viewpoint, with subjects maintaining a consistent facing position towards the camera throughout their performances. Severe post-processing techniques were subsequently applied to eliminate the background, enhancing the dataset's clarity and focus. It is noteworthy that for actions necessitating the use of one arm or one leg, actors were uniformly instructed to perform them using their right arm or leg, improving consistency and promoting accurate analysis.However, this design introduces potential biases in asymmetric action modeling and limits multi-viewpoint generalizability, unlike datasets like UCF101 [33]. While the dataset's controlled environment and 3D skeletal data (derived from depth sensing) make it suitable for evaluating sensor-based motion analysis, its reliance on a single sensor contrasts with NTU RGB+D [145], which employs multi-camera setups for higher joint estimation accuracy. The lack of explicit annotation protocols (e.g., inter-annotator agreement metrics) and limited environmental complexity (e.g., no occlusions or dynamic backgrounds) may restrict its applicability to real-world scenarios. Action labels were manually defined by researchers without documented validation, unlike HMDB51 [163], which incorporates crowdsourced annotations.

3.2.4 HMDB51

The HMDB51 dataset[163] constitutes an essential repository of realistic videos derived from a wide array of media sources, including motion pictures and web-based videos. It encompasses a total of 6,766 video segments, mainly categorized into 51 unique action classes, with each class containing no fewer than 101 clips. In addition, Videos range in resolution (320×240–1280×720) and frame rate (15–30 fps), with minimal preprocessing (e.g., no background removal).For the purpose of evaluation, the dataset adopts an innovative methodology that involves dividing the data into three distinct training or testing splits. Within each of these splits, 70 clips per action class are allocated for training purposes, while the remaining 30 clips are reserved for testing. This split strategy mitigates overfitting risks compared to fixed partitions, though the limited class

diversity (51 vs. 101/1,000+ in UCF101[33]/Kinetics[102]) and short video durations (avg. 3 seconds) may hinder temporal modeling benchmarks.

The action categories in the HMDB51 dataset could be systematically classified into five types: 1) general facial expressions: smiles, laughter, chewing, and talking; 2) facial actions involving object manipulation: smoking, eating and drinking; 3) general body movements: this group comprises a diverse array of movements, ranging from cartwheels and clapping hands to climbing, ascending stairs, diving, falling, backhand flips, hand-stands, jumping, pull-ups, push-ups, running, sitting, somersaults, standing, turning, walking, and waving; 4) body movements with object interaction: brushing hair, catching, drawing sword, dribbling, golf , hitting something, kicking ball, picking, pouring, pushing something, riding bike, riding horse, shooting ball, shooting bow, shooting gun, swinging baseball bat, sword exercise and throwing; 5) body movements for human interaction: fencing, hugging, kicking someone, kissing, punching, shaking hands and sword fight.However, the dataset's fine-grained action definitions (e.g., distinguishing "sword exercise" from "sword fight") demand precise spatio-temporal modeling, unlike coarser categories in UCF101 [33].

3.2.5 UCF101

The UCF101 dataset[33], an extensive extension of the UCF50, encompasses a meticulously curated collection of 13,320 video clips, mainly divided into 101 unique classes. These classes are further segmented into five distinct categories: body motion, human–human interactions, human–object interactions, playing musical instruments, and sports. The aggregate duration of these video clips surpasses 27 h, offering a substantial amount of data for analysis. All videos within this dataset are sourced exclusively from YouTube and exhibit a uniform frame rate of 25 frames per second (FPS), coupled with a resolution of 320×240 pixels, ensuring consistency and quality. The dataset undergoes minimal preprocessing (e.g., no background removal) to preserve raw visual diversity, but its reliance on YouTube sources introduces variability in lighting, camera angles, and motion scales. Action labels are manually curated by researchers with explicit class definitions (e.g., distinguishing "golf swing" from "tennis serve"), though no documented inter-annotator agreement metrics are provided, contrasting with HMDB51 [163], which employs crowdsourced annotations with over 80% consensus. This dataset is specifically designed and intended for rigorous academic use, catering to the needs of researchers and scholars in the field. Compared to HMDB51 [163], which emphasizes fine-grained human interactions (e.g., fencing, handshakes), UCF101 focuses on broader, sports-related actions (e.g., running, swimming) with clearer motion patterns. However, unlike Kinetics [?] (10,000+ classes, real-world diversity) or NTU RGB+D [145] (3D skeletal data, multi-view scenarios), UCF101 lacks 3D joint information and struggles with dynamic occlusions due to its 2D video-only format. Fig. 11 shows the typical action classes of UCF101. While its structured class hierarchy and consistent resolution make it ideal for benchmarking action detection algorithms, the absence of depth data and limited environmental complexity (e.g., no multi-person interactions) may restrict its applicability to real-world scenarios requiring robust geometric reasoning.

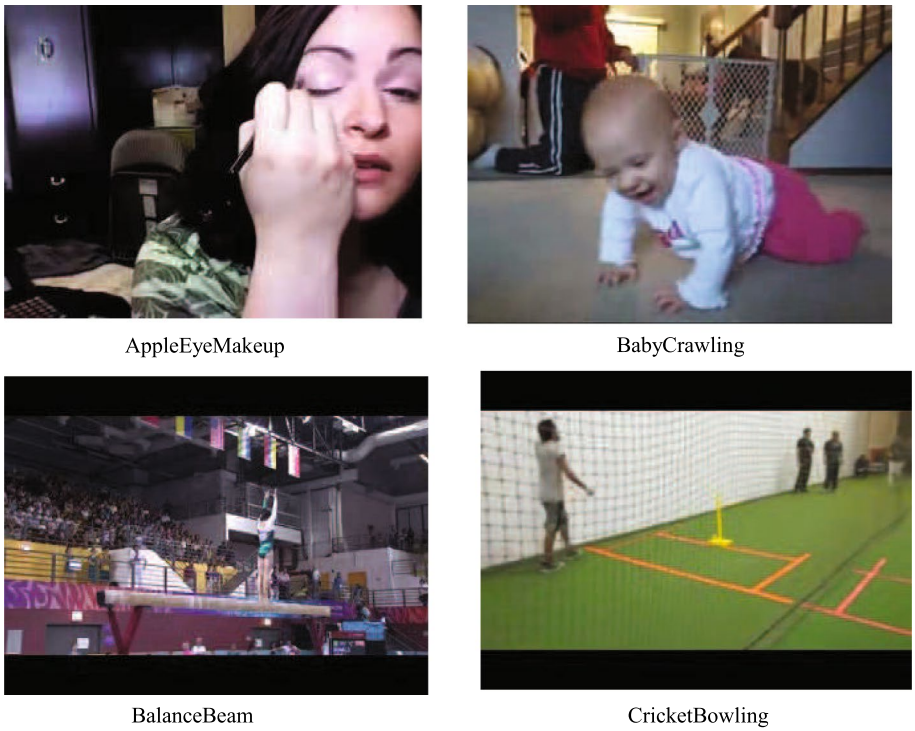

Fig. 11 The typical action classes of UCF101

3.2.6 Sports-1M

The Sports-1M dataset[166] is a comprehensive collection of over one million videos sourced from YouTube, accessible via the URLs provided by the dataset creators. As of 2016, despite the fact that approximately 7 percent of the videos in the dataset had been re-moved by their respective uploaders after the dataset was compiled, it still boasts a robust collection of over one million videos.These videos exhibit a wide range of resolutions (e.g., 320×240 to 1280×720) and frame rates (15–30 FPS), reflecting the uncurated nature of YouTube content, but also introducing variability in visual quality and motion clarity. And they are precisely categorized into 487 sports-related classes, with each class containing a substantial range of 1,000 to 3,000 videos. However, the automated annotation process, which relies on the YouTube Topics API to extract textual metadata (e.g., tags, descriptions), may result in label inaccuracies due to ambiguities in user-generated content (e.g., mislabeled "soccer" as "football" in non-English contexts).The annotation process for these videos is highly automated, utilizing the YouTube Topics API to analyze the textual metadata associated with the videos, such as tags and descriptions. This results in the videos being accurately annotated with one of the 487 sports categories.While around 5 percent of the videos are annotated with multiple classes, this reflects both the inherent complexity of sports content (e.g., a video of "tennis match" might also involve "running" or "sprinting") and potential overlaps in the hierarchical class structure. It is worth noting that around 5 percent of the videos in the dataset have been annotated with multiple classes, reflecting the complexity and diversity of the sports content.Compared to UCF101 [33] (101 classes, 27 hours total duration) and Kinetics [102] (1,000+ classes, real-world diversity), Sports-1M offers

a massive scale (1,000,000 videos) but lacks fine-grained action definitions and 3D joint data, limiting its utility for tasks requiring precise spatio-temporal modeling or geometric reasoning (e.g., NTU RGB+D [145]). This dataset is specifically intended for academic use, providing researchers with a valuable resource for conducting in-depth analyses and gaining insights into the world of sports through the lens of video data. Its primary strength lies in large-scale pretraining for sports-specific models, though its reliance on automated annotations and lack of manual validation may hinder performance in downstream tasks requiring high-precision benchmarks.

3.2.7 N-UCLA

The N-UCLA dataset [142], whose full name is Northwestern-UCLA Multiview 3D event dataset, contains RGB, depth and skeleton data captured synchronously via three spatially calibrated Kinect cameras, ensuring temporal alignment among the three modalities. This dataset provides high-resolution RGB videos (640×480 pixels, 30 FPS) and depth maps (320×240 pixels, 30 FPS), with 3D skeleton data derived from joint tracking using Microsoft Kinect v2. The synchronized multi-modal data enable robust modeling of geometric relationships and motion patterns across views.This multi-view HAR dataset is composed of 10 action classes: pick up with one hand, pick up with two hands, drop trash, walk around, sit down, stand up, donning, doffing, throw and carry. Each action is annotated with precise 3D joint coordinates (21 joints) for all actors, allowing for explicit modeling of body part configurations and spatial dependencies. The dataset's cross-view design (three fixed camera angles: 0°, 90°, and 180°) ensures consistent coverage of actions from different perspectives, though the limited number of views (3 vs. 10+ in Kinetics [102]) may restrict generalization to arbitrary viewpoints. And each of the actions is performed by 10 actors. The dataset's structured setup (10 actors × 10 actions × 3 views) facilitates benchmarking for cross-view recognition, with extensive experiments demonstrating its utility for tasks like pose estimation and view interpolation. However, compared to large-scale datasets like Kinetics or UCF101 [33], its smaller class diversity (10 vs. 101/1,000+ classes) and controlled environment limit its applicability to real-world scenarios with complex backgrounds or unstructured motion.

3.2.8 ActivityNet

Activity dataset [100] collects 200 distinct types of activities and a total of 849 hours of videos from YouTube. As of now, ActivityNet is one of the largest benchmarks for temporal activity detection in terms of both the quantity of activity categories and the number of videos, rendering the task exceptionally challenging. Version 1.3 of this dataset consists of a total of 19,994 untrimmed videos and is divided into three mutually exclusive subsets, namely training, validation, and testing, in a ratio of 2:1:1. This untrimmed format (avg. 25.6 minutes per video) introduces complex background clutter and overlapping activities, making it significantly harder than trimmed datasets like Kinetics [?] (short clips, 1,000+ classes) or Sports-1M [166] (automated annotations, 487 classes).Each activity category has 137 untrimmed videos and each video contains 1.41 activities that are annotated with temporal boundaries.The annotations are crowdsourced via Amazon Mechanical Turk with explicit guidelines, achieving over 85% inter-annotator agreement for start/end timestamps, unlike manually curated datasets

like Multiview Action3D [142] that rely on 3D skeleton data for geometric modeling. Compared to Sports-1M, ActivityNet's scale (200 vs. 487 classes) and focus on precise temporal boundary detection enable robustness testing for real-world scenarios, though its lack of 3D joint data limit geometric reasoning tasks compared to Multiview Action3D.

3.2.9 NTU RGB+D

The NTU RGB+D dataset [145] is a vital and widely used dataset for 3D human activity analysis in the field human action recognition, which was introduced by researchers from institutions such as Nanyang Technological University. It is a large-scale human action dataset which contains more than 56,000 multi-view RGB-D videos with 4 million frames that include 60 actions performed by 40 subjects. The actions in the dataset could be generally divided into three main categories: 40 daily actions (such as drinking, eating and reading), 9 health-related actions (such as sneezing, staggering and falling down), and 11 mutual actions (like punching, kicking and hugging).This structured class hierarchy enables fine-grained analysis of both solitary and social behaviors, distinguishing it from broader datasets like Sports-1M [166] (487 sports categories) or ActivityNet [100] (200 diverse activities).

The actions were captured using three cameras with different horizontal imaging perspectives, namely, −45°, 0°, and 45°. This multi-view capture provides rich information for analyzing human actions from different perspectives.Compared to N-UCLA [142] (10 actions, 3 fixed views), NTU RGB+D's larger scale (60 actions) and varied camera angles (3 vs. 10+ in Kinetics [102]) enhance robustness for cross-view recognition tasks. The dataset offers multi-modality information for each action sample, including depth maps, 3D skeleton joint positions, RGB frames, and infrared sequences. The inclusion of 3D skeleton data (25 joints per frame) allows explicit modeling of body part configurations, unlike datasets like ActivityNet that rely solely on 2D video inputs.The depth maps and infrared videos have a resolution of 512×424, while the RGB videos have a high resolution of 1920×1080. And the most competitive advantage of it is that this dataset releases two validation protocols namely cross subject (CS) and cross view (CV).These protocols facilitate benchmarking for both intra-subject and cross-view generalization, addressing limitations in datasets like Sports-1M (automated annotations) or ActivityNet (crowdsourced temporal boundaries).Fig. 12 shows the different modalities contained in the NTU RGB+D dataset.

3.2.10 THU-READ

The THU-READ dataset [176], which could also be called Tsinghua University RGB-D Egocentric Action Dataset. It is designed for action recognition in egocentric

(a) RGB (b) RGB+joints (c) depth (d) depth+joints (e) IR

Fig. 12 Different modalities of NTU RGB+D dataset[145]

videos, which means the videos were captured from the first-person perspective. This dataset contains 40 action classes that are "all-about-hand", focusing on hand-related actions. It was created by 8 subjects(6 males and 2 females with a height range from 162 cm to 185 cm). While collecting the data, an RGB-D sensor was mounted on a helmet and placed on the subject's head. The camera was kept in the same direction as the subject's eyesight to simulate real conditions for acquiring egocentric action videos. Compared to large-scale datasets like Kinetics [102] (1,000+ classes, short clips) or Sports-1M [166] (487 sports categories), THU-READ offers a more structured class hierarchy focused on fine-grained hand-centric actions, making it particularly suitable for tasks like grasp understanding or social interaction analysis. And to balance the data distribution, each action class was performed by each subject for the same number . The inclusion of 3D skeleton data (25 joints per frame) enables explicit modeling of hand-object spatial relationships, unlike datasets like Multiview Action3D [142] (10 actions, 3 fixed views) that lack detailed hand-specific annotations. Eventually, the dataset obtained 1920 video clips, calculated as:

$$8 \times (\text{subjects}) \times 2 \times (\text{modalities - RGB and depth}) \times 40 \times (\text{classes}) \times 3 \times (\text{times}) \tag{1}$$

3.2.11 Kinetics-400

The Kinetics-400 [102] dataset could be regarded as the successor to the HMDB51 [163] and UCF101 [33]. However, both of them have insufficient size and are lack of sufficient variation for training and evaluating the current generation of deep learning-based human action classification models.This limitation is addressed by Kinetics-400, which expands the number of classes from 10/51 to 400, offering a much larger and more diverse set of actions compared to earlier datasets like Sports-1M [166] (487 sports categories) or Multiview Action3D [142] (10 actions). And the increase of the number of classes then increases from 10/51 to 400. And each action class is represented by at least 400 video clips. Each clip is approximately 10 seconds in duration and sourced from a unique YouTube video. This design ensures high inter-class variability, as each clip captures distinct scenarios, lighting conditions, and actor demographics, unlike the more controlled settings of Multiview Action3D or the automated annotations of Sports-1M.The actions are centered on human activities and cover a wide range of classes, including human–object interactions (e.g., playing musical instruments) and human–human interactions (e.g., shaking hands), suitable for academic research and analysis.Compared to ActivityNet [100] (200 classes, 849 hours), Kinetics-400 provides more granular class definitions (400 vs. 200) and shorter, more focused video clips (10s vs. untrimmed long videos), which aligns better with the needs of fine-grained action recognition tasks. However, unlike Multiview Action3D, Kinetics-400 lacks explicit 3D skeletal annotations, limiting its utility for geometric reasoning tasks. Additionally, the use of crowdsourced temporal boundaries in ActivityNet contrasts with Kinetics-400's automated labeling via YouTube metadata, which may introduce variability in annotation quality. Fig. 13 shows four typical action frames from the Kinetics-400 dataset.

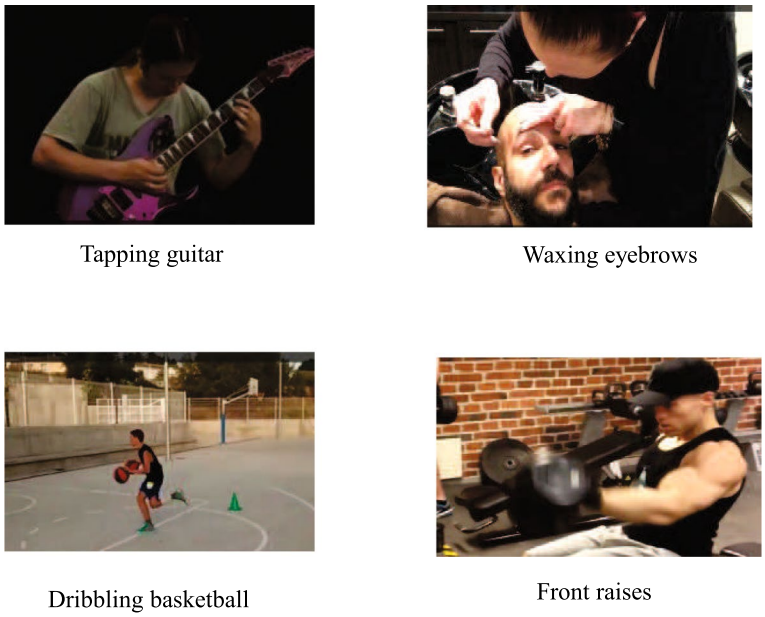

Fig. 13 The typical action classes of Kinetics-400 dataset[102]

3.2.12 Something-something

Like the Kinetics datasets, the something-something datasets [179] has more than one version, Something-Something v1 and something-something v2. To be honest, the v2 is used more frequently in today's research. The something-something dataset v1 and v2 are a large collection of labeled video clips which show human performing pre-defined basic actions with everyday objects.Unlike broader datasets like Kinetics [102] (1,000+ classes, general activities) or ActivityNet [100] (200 classes, untrimmed long videos), something-something focuses on highly specific, fine-grained actions involving object manipulation (e.g., "put something in front of something" vs. "take something out from something"), making it ideal for tasks like robotic imitation learning or action parsing. The v1 version, compiled by a substantial number of crowd workers, facilitates the development of fine-grained understanding in machine learning models regarding basic actions occurring in the physical world. It comprises 108,499 videos, allocated as follows: 86,017 for the training set, 11,522 for the validation set, and 10,960 for the test set. Additionally, it encompasses a total of 174 distinct labels.This class hierarchy is more structured than Sports-1M [166] (487 sports categories) or ActivityNet (200 diverse activities), as it explicitly defines actions based on spatial–object relationships rather than broad activity types. As for the v2, the total number of videos is different from the v1, since it contains 220,847 videos, with 168,913 in the training set, 24,777 in the validation set and 27,157 in the test set.

Both versions emphasize human–object interaction semantics (e.g., "move something from one container to another"), contrasting with Multiview Action3D [142] (10 actions, 3D skeleton annotations) or NTU RGB+D [36] (60 actions, 3D joints), which focus on geometric reasoning but lack explicit object-centric labeling. The dataset's reliance on crowdsourced annotations (via Amazon Mechanical Turk) ensures high inter-annotator

agreement (over 85%) for action boundaries, surpassing the automated metadata-driven labeling of Sports-1M or ActivityNet.

3.2.13 NTU RGB+D 120

The NTU RGB+D 120 dataset [181] could be regarded as the successor of the NTU RGB+D dataset, which is collected from 106 distinct subjects and contains more than 114 thousand video samples and 8 million frames. This dataset contains 120 different action classes including daily, mutual, and health-related activities. This represents a significant increase from the original 60-class NTU RGB+D, and it offers a more comprehensive action hierarchy compared to large-scale datasets like Kinetics [102] (1,000+ classes) or Sports-1M [166] (487 sports categories), which prioritize broad activity diversity over structured classification.

The differences between NTU RGB+D and NTU RGB+D 120 are shown in the Table 7, which could show readers the details of the contrasts of the two datasets. Compared to N-UCLA [142] (10 actions, 3 fixed views), NTU RGB+D 120's larger scale (120 actions) and multi-view capture enhance robustness for cross-subject and cross-view generalization, addressing limitations in datasets like ActivityNet [100] (200 classes, untrimmed long videos).The dataset retains the multi-modality information of its predecessor, including RGB frames, depth maps, 3D skeleton joint positions, and infrared sequences, with RGB videos at high resolution (1920×1080) and depth/infrared at 512×424. This synchronized multi-modality setup enables precise spatio-temporal analysis, surpassing the 2D-only format of ActivityNet and the automated metadata-driven labeling of Sports-1M.

3.2.14 ETRI-Activity3D

The ETRI-Activity3D [182] dataset, designed for robots to recognize daily activities of the elderly from a robot's perspective, contains 112,620 samples with RGB videos, depth maps, and skeleton sequences. This dataset covers 55 daily action categories, carefully selected through meticulous observations of the daily lives of 53 elderly individuals aged over 70.The categories contains both common daily actions (e.g., eating, cleaning, reading) and human–robot interaction-specific actions(e.g., waving, beckoning, pointing), with 5 mutual actions. This structured class hierarchy, which combines general daily activities and robot-specific interactions, distinguishes it from broader datasets like Kinetics [102] (1,000+ classes) or ActivityNet [100] (200 classes), which lack explicit

Table 7 The comparison between NTU RGB+D 120 and NTU RGB+D

Metric	NTU RGB+D	NTU RGB+D120
Number of action classes	60	120
Sample size	56,880	114,480
3D skeletal data	25D coordinates of 3 body joints per frame	3D coordinates of 25 human joints per frame
Data collection devices	Kinect v2 sensor	Kinect v2 sensor + additional calibration tools
Video resolution	1920×1080	1920×1080 (with higher frame rate)
Annotation method	Manual labeling	Semi-automated labeling with validation
Participant count	40	106

focus on human–robot interaction scenarios.Moreover, the data were captured in an apartment setting that mirrors typical elderly living conditions. Compared to N-UCLA [142] (10 actions, 3 fixed views) or NTU RGB+D [181] (60 actions, lab-controlled environments), ETRI-Activity3D emphasizes real-world ecological validity by simulating natural aging environments (e.g., kitchen, living room) with realistic lighting and clutter. The sensors were positioned at 70 cm and 120cm heights with distances varying from 1.5-3.5m, capturing actions from four platforms to enable multi-view analysis. And the subjects, including 50 elderly and 50 young adults, performed actions in their own styles, across different periods during one day. In summary, ETRI-Activity3D bridges the gap between academic research and real-world robotic applications by prioritizing ecological validity, scalability, and demographic representation.

3.2.15 UAV–human

The UAV–human dataset [183] stands as a comprehensive and invaluable resource for exploring human behavior through the utilization of unmanned aerial vehicles (UAVs). It comprises an extensive collection of 67,428 multi-modal video sequences, emphasizing 119 subjects for action recognition. Besides, it incorporates 22,476 frames specifically designated for pose estimation, 41,290 frames with 1,144 distinct identities for person re-identification, and 22,263 frames focused on attribute recognition.Unlike general-purpose datasets like Kinetics [102] (1,000+ classes) or ActivityNet [100] (200 classes, untrimmed videos), UAV–human integrates multiple tasks (action, pose, re-ID, attributes) under a unified UAV perspective, enabling cross-task analysis. This extensive dataset was precisely gathered over a span of three months by a UAV operating across a variety of urban and rural scenes, during both daytime and nighttime hours. Consequently, it summarizes a diverse range of variations, including subject types, backgrounds, lighting conditions, weather phenomena, occlusions, camera movements, and UAV flight attitudes. This contrasts with Sports-1M [166] (487 sports categories, single-view metadata) and Multiview Action3D [142] (10 actions, fixed views), by emphasizing real-world complexity and multi-view dynamics.The UAV–human dataset acts as an essential tool for UAV-based research endeavors aimed at understanding human behavior, encompassing a wide array of tasks such as action recognition, pose estimation, person re-identification, and attribute recognition.

3.2.16 HA4M

3.2.17 Human-art

3.2.18 WiFall

The WiFall dataset [185] is a Wi-Fi sensing dataset designed for cross-domain human action recognition and fall detection tasks. It contains Channel State Information (CSI) data with a sampling rate of 100 HZ, 1 antenna and 52 subcarriers captured by an ESP32-S3 receiver and a commercial Wi-Fi router. This dataset includes 10 volunteers, performing five action categories: walking, jumping, sitting, standing up and falling. Unlike general-purpose datasets like Kinetics [102] (1,000+ classes) or Sports-1M [166] (487 sports categories), WiFall focuses on critical safety-related actions (e.g., fall

detection) using low-cost Wi-Fi hardware, making it ideal for assistive technologies in elderly care or smart homes.Compared to ActivityNet [100] (200 classes, untrimmed long videos), WiFall provides short, structured video clips (1s duration) with precise temporal boundaries, enabling robust training for fine-grained action recognition. Its use of CSI data (52 subcarriers, 1 antenna) also introduces unique spatio-temporal patterns distinct from RGB/depth-based datasets like N-UCLA [142], which lack Wi-Fi-specific modality.The dataset's cross-domain design (e.g., varying environments, lighting, and occlusions) ensures applicability to real-world deployment challenges, such as domain shift in wireless sensing scenarios.

3.3 Continuous datasets

Continuous datasets pertain to those datasets in which each video sequence may contain one or multiple actions/gestures, with the boundaries delineating distinct motion classes remaining undisclosed. These datasets are predominantly employed for tasks such as action detection, localization, and real-time action prediction. The following subsections would introduce some typical ones.

3.3.1 ChaLearn2014

ChaLearn 2014 Multimodal Gesture Recognition [168] is a comprehensive collection of multi-modal data which are acquired using a Kinect v1 sensor(http://gesture.chalearn.org/2014-looking-at-people-challenge.) It covers RGB images, depth maps, skeleton data, and audio recordings. Throughout the dataset, a solitary individual is filmed in front of the camera, executing natural Italian gestures that are commonly used in communication. Each gesture in the dataset is precisely annotated with both the gesture class label and the precise starting and ending frames. The dataset comprises approximately 14,000 manually labeled gesture instances, occurring within continuous video sequences. These gestures belong to a vocabulary of 20 distinct Italian gesture categories.

In total, there are 1,720,800 labeled frames distributed across 13,858 video fragments, each lasting approximately 1 to 2 min and sampled at a rate of 20 Hz. The gestures were performed by 27 different individuals under a variety of conditions, including variations in clothing, positioning, backgrounds, and lighting. This dataset provides a rich resource for research in multi-modal gesture recognition.

3.3.2 Charades

Charades dataset [172] is a comprehensive, large-scale dataset, primarily focusing on documenting everyday household activities through the innovative Hollywood in Homes methodology. The name of the dataset is inspired by a renowned American word-guessing game where participants mime phrases for others to decipher. Along similar lines, they recruited hundreds of participants from Amazon Mechanical Turk to enact the paragraphs, who also contributed annotations for action classification, localization, and video descriptions.The inaugural publicly released version of the Charades dataset boasts 9848 videos, each averaging 30.1 seconds in length. These videos are mainly divided into 7,985 training videos and 1,863 test videos.Unlike general-purpose datasets like Kinetics [102] (1000+ classes) or ActivityNet [100] (200 classes, untrimmed videos),

ChaLearn focuses on fine-grained, context-specific gestures with precise frame-level annotations, making it ideal for tasks like sign language recognition or human–computer interaction. The dataset spans 15 diverse types of indoor scenes, encompassing interactions with 46 object classes. It boasts a rich vocabulary of 30 verbs, resulting in the delineation of 157 distinct action classes. What's more, the dataset encompasses 66,500 temporally localized actions, averaging 12.8 seconds in duration. These actions were recorded by 267 individuals spanning three continents.

3.3.3 PKU-MMD

PKU-MMD [175] is a dataset specifically using in the fields of long continuous sequence action detection and multi-modality action analysis. This dataset is captured with the Kinect v2 sensor, which has the capability to synchronously collect color images, depth images, infrared sequences, and human skeleton joints. Within this dataset, it has amassed over 1,000 extended action sequences, each spanning approximately 3 to 4 minutes (captured at a frame rate of 30 FPS) and encompassing approximately 20 action instances. The comprehensive scale of this dataset covers 5,312,580 frames, which translates to 3,000 minutes of video content, featuring more than 20,000 temporally localized actions. PKU-MMD selected a total of 51 action classes and divided them into two categories: 41 daily actions (such as drinking, waving hand, putting on glasses, etc.) and 10 interaction actions (such as hugging, shaking hands, etc.).This structured class hierarchy, combining routine and social interactions, contrasts with general-purpose datasets like Kinetics [102] (1,000+ classes) or N-UCLA [142] (10 actions), which lack explicit focus on daily-life and human–human interaction tasks. To assemble this extensive dataset, they engaged 66 unique individuals in our data acquisition process. Each participant contributed to the recording of four daily action videos and two interactive action videos. The age range of the participants was between 18 and 40 years, thereby ensuring a diverse and comprehensive sample suitable for human action recognition.This balanced age distribution and real-world activity design enhance demographic and contextual diversity, addressing limitations in datasets like Sports-1M [166] (487 sports categories, metadata-driven annotations) or ActivityNet's untrimmed format. As for the data acquisition process, each participant contributed to the recording of four daily action videos and two interactive action videos. Table 8 shows the SOTA algorithms for the PKU-MMD dataset.

4 Video action recognition combined with methods in other fields

From the introduction in the section2, it is easy to say that the video action recognition has a huge progress due to the methods mentioned in the section2. Although the pure methods for video action recognition could work, they still could not overcome some tough problems. If we could use the methods combined with other fields in computer vision, many tough problems could be solved soon. Table 9 shows the comparison for the methods that combined with object detection and object tracking.

4.1 Combined with object detection

The integration of object detection and action recognition offers several significant advantages in the field of computer vision. By combining object detection and action

Table 8 SOTA performance comparison on PKU-MMD [175].

Method	Year	Metric (%)	Extra training data	Core features
DVANet [170]	2024	95.8(CS) 95.2(CV)	×	Transformer-based disentanglement with contrastive losses for robust multi-view action recognition.
TSMF [230]	2021	95.8(CS) 97.8(CV)	✓	Teacher–student multimodal fusion (TSMF) fuses RGB and skeleton modalities
EPAM-Net [218]	2025	96.2(CS) 98.4 (CV)	×	Efficient pose-driven attention-guided multimodal network with 6.2–9.9x FLOPs and 9–9.6x parameter reductions.
MMNet [217]	2023	97.4(CS) 98.6(CV)	✓	A model-based multimodal fusion for RGB-D action recognition using ST-GCN attention transfer.
DSCNet [174]	2023	97.4(CS) 98.8(CV)	×	RGB + Pose, with transformer, achieves SOTA performance on PKU-MMD.

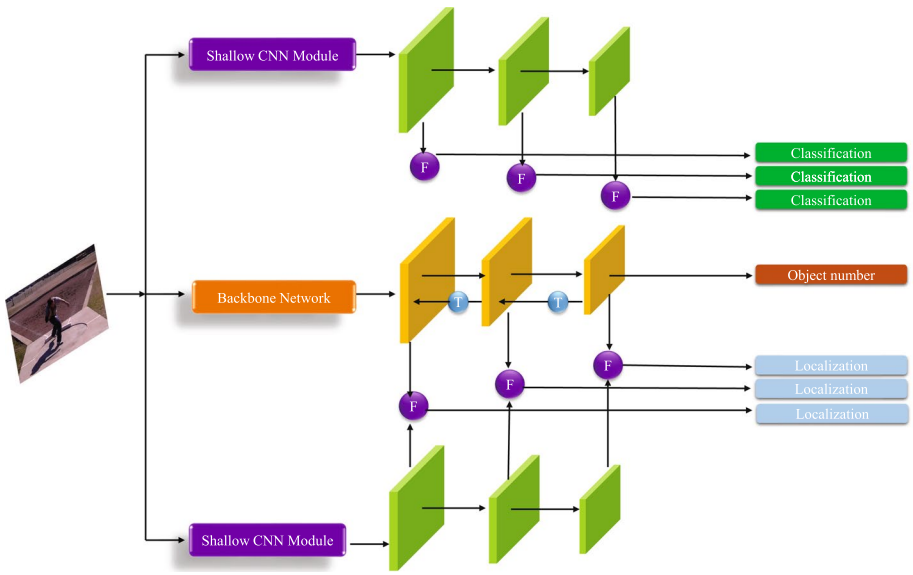

Fig. 14 The overall architecture of the multi-object behavior recognition framework

recognition, systems can gain a deeper understanding of the visual scene. Object detection identifies the presence and location of various objects, while action recognition interprets the behaviors and interactions associated with these objects. Elujide et al. [231] proposed a deep learning network combined human action recognition and object detection for CSI-based multiple human activity recognition. This network could precisely determine the temporal occurrence of each activity within a stream, accurately classify the action labels along with their corresponding confidence scores, and further generate a unique segmentation mask to distinguish between various instances of the same action. What's more, Dang et al. [232] presented a shallow CNN module, using the idea of object detection, to complete the multi-object action recognition in dense crowd. Besides, they introduced a feature fusion mechanism aiming at enhancing the detection performance of small objects through the acquisition of multi-scale high-level semantic information, as shown in Fig. 14. Moreover, Du et al. [233] proposed a two-stage framework for interaction recognition through technical innovations, which

contains deformable convolutional layers with learnable sampling offsets enabling adaptive receptive field adjustment for complex scenes, and an optimized detection network combining attention mechanisms and pre-activated ResNet blocks(BN-ReLU-Conv) to enhance localization precision and regularization. In addition, Hsiao et al. [234] combines YOLO and DG-STGCN, which highly enhanced the accuracy of action recognition. Table 9 shows the experimental results and other essential details for the methods mentioned in this subsection.

4.2 Combined with object tracking

Human tracking and action recognition stands as a pivotal research area within the realm of computer vision, which is also commonly referred to as real-time video processing. This domain harbors immense potential for a wide array of applications spanning various fields, notably including augmented reality, human–machine interaction, and advanced driver assistance systems, among numerous others.

However, the integration of action recognition with object tracking is critical for achieving robust and context-aware behavior analysis. While tracking ensures continuous localization of targets (e.g., individuals or vehicles) over time, action recognition provides semantic understanding of their behaviors (e.g., "falling", "running", or "gesturing"). Together, they enable systems to not only locate entities but also interpret their actions in real-world scenarios. For instance, in augmented reality, precise tracking of a user's position combined with gesture recognition allows seamless interaction with virtual objects, while in autonomous driving, this synergy enhances safety by detecting pedestrian intentions (e.g., "crossing" or "hesitating"). Despite these benefits, challenges such as occlusion handling, multi-target association, and real-time performance still require further investigation to unlock the full potential of this integrated approach. Table 10 gives a comprehensive summary for the methods mentioned in this subsection.

4.2.1 Improved action recognition based on specific tracking model

Duan et al. [235] proposed a method for recognizing irregular behavior in laboratory personnel by utilizing an improved DeepSORT [236] algorithm catered to the specific characteristics of a chemical laboratory setting. The methodology initially involves the extraction of skeletal keypoints from laboratory personnel, using the Lightweight OpenPose algorithm for the purpose of individual localization. Subsequently, the refined DeepSORT algorithm is employed to track human targets and ascertain the positions of pertinent objects. Eventually, an SKPT-LSTM network is taken to integrate the tracking data, thereby enabling action recognition. Besides, Zhou et al. [237] presented a novel human action recognition method, which entailed the extraction of keyframes via an enhanced density clustering technique, with the acquisition of spatio-temporal context information being assisted by a Context-Guided Bidirectional Long Short-Term Memory (BiLSTM) network. Moreover, Chang et al. [238] presented a real-time system for basketball player action recognition and tracking, aiming to enable precise court strategy analysis for coaches and players.The proposed framework integrates YOLOv8 for real-time object detection, BoT-SORT for multi-player tracking, and an R(2+1)D CNN with ResNet50 backbone for spatio-temporal feature extraction and action classification.This

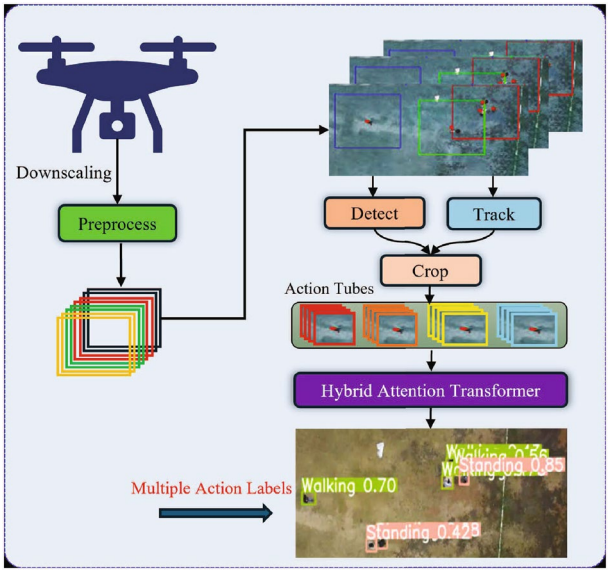

Fig. 15 The whole architecture of HAT [240]. Figure from [240]

Table 9 Comparison for methods combined with object detection

Methods	Detection algorithms	HAR algorithms	Metric(%)	Key contribution
Elujide et al. [231]	Mask R-CNN	CSI-based network	93.8 WiFI CSI [231]	First real-time object detection framework for WiFi CSI-based HAR.
Dang et al. [232]	YOLOv7	Shallow CNN	mAP50 88.72 SCB [232]	A novel shallow CNNs module to assist localization and classification parallel branch structures
Du et al. [233]	Faster RCNN	Object detection + Action prediction	mAP 67.2 V-COCO [243]	It divides the action recognition process into target detection and action prediction.
Hsiao et al. [234]	YOLOv8	DG-STGCN	65(without YOLO) 92(with YOLO)[1]	Heterogeneous fusion architecture Dynamic decision framework

[1] The result is validated on assembly process simulation dataset made by the authors.

work highlights the effectiveness of integrating deep learning-based detection, tracking, and spatio-temporal feature learning for real-world sports analytics.

4.2.2 Action tracking and recognition with integrated attention mechanism

Choi et al. [239] are the first researchers who attended the filed of combining object tracking and action recognition. They proposed a unified and discriminative framework aiming at concurrently tracking multiple individuals and inferring their collective activities. Instead of tackling these two issues in isolation, their model was anchored on the insight that there existed a significant correlation among an individual's motion, their activity, and the movements and activities of other nearby individuals. Instead of directly connecting the resolutions of these two challenges, they proposed a hierarchical

Table 10 Comparison for methods combined with object tracking.

Methods	Tracking algorithms	HAR algorithms	Metric[1](%)	Key contribution
Duan et al. [235]	Improved DeepSORT	SKPT-LSTM	Accuracy 87.99 F1 Score:0.8654	Developed SKPT-LSTM to fuse skeleton and object position sequences for action analysis.
Zhou et al. [237]	Improved density clustering	Context-guided BiLSTM	78.1 (JHMDB [101])	Proposes a context-guided BiLSTM model with residual connections and classification parallel branch structures
Chang et al. [238]	BOT-SORT	R(2+1)D with ResNet50	Action recognition: 85 YOLOv8 mAP50: 94.32	Employ R(2+1)D CNN to separate spatial (2D) and temporal (1D) convolutions
Hsiao et al. [239]	Min-cost flow Branch-and-Bound	Hierarchical graphical model	83.0(Action recognition) 82.78(Object tracking)	Unified framework for joint multi-target tracking and collective activity recognition.
Drone HAT [240]	DeepSORT	HAT	60.76 Okutama [243])	Multi-Label HAR Framework DeepSORT integration
RETA [241]	Integrated detection/tracking	DC-RCNN	94.89 83.49[2]	Joint detection/tracking DC-RCNN with CTC 4D Radar Utilization
Wang et al. [242]	Centroid tracking algorithm	YOLOv5 + C3D	Carrying luggage: 60 Not carrying luggage: 83	Multi-target action recognition Centroid tracking for sequences

[1] Unless otherwise stated, the results are validated on the dataset made by the authors.

[2] Continuous recognition accuracy: 94.89% (normal conditions) and adverse conditions: 83.49%.

categorization of activity types that delineated a natural progression from an individual's motion to the collective activity covering the entire group. Their model is endowed with the ability to concurrently track numerous individuals, recognize individual activities (atomic activities), differentiate interactions between pairs of individuals (interaction activities), and ultimately, grasp the behavioral patterns of groups of people (collective activities). Inspired by the work of Choi et al.[239], Khan et al. [240] proposed a framework for multi-label action detection, object tracking and action recognition that leveraged a Hybrid Attention Vision Transformer(HAT) to enhance the efficiency of recurrent action recognition. Furthermore, within the transformer block dedicated to action recognition, a multi-scale, multi-granularity module is incorporated to extract pertinent features in a non-redundant manner. The whole architecture of HAT is shown in Fig. 15.

4.2.3 Object detection, object tracking, and action recognition integrated

Zhang et al. [241] introduced an end-to-end joint tracking and action estimation(RETA) system based on 4D auto-motive radar, which specifically focused on identifying pedestrian activity within complex real-world scenarios. To differentiate activities of varying durations, a decomposed CRCNN was proposed, enhancing fused spatio-temporal

feature extraction. The labor-intensive presegmentation was bypassed with a connectionist temporal classification algorithm. Eventually, the RETA system is suitable for real-world end-to-end perception applications. Moreover, Wang et al. [242] introduced a general algorithm for the combination of object tracking and human action recognition when it comes to large-size pixel video and action video of multiple people with different actions at the same time. The proposed network combines YOLOV5 object detection network, centroid tracking algorithm and C3D video action recognition network. As a result, it could recognize the behaviors of multiple people in a single video.

5 The applications of human action recognition in different areas

Human action recognition applies across various fields, offering research and development challenges and opportunities. It aids in-depth analysis in sports, medical rehabilitation and traffic safety by providing quantifiable human behavior data. These applications emphasize its significance and versatility in modern technology.

5.1 The applications in sports

In recent years, researchers in the communities of computer vision and sports pay much attention to sports video analysis. And in the section, we would categorize the applications into the following subsections.

5.1.1 Video judge

In the past few decades, the video-based match judge has been widely applied in the sports games, where most of them select human action recognition as the primary assistant to release the fairest decision from the referee. In particular, Nekoui et al.[244] proposed a virtual refereeing network to valid the execution of a diving performance. This evaluation would rely on both visual cues and the sequence of body joints in the action video. To address the unusual body movements in such situations, they introduced ExPose: annotated dataset of Extreme Poses. In their network, they adopted ST-GCN [130] and HRNet [245] to assess the difficulty of the performance based on the extracted joints sequence. Eventually, the final score for the performance would be calculated by multiplying the execution score by the difficulty score. Instead of judging the performance of the athletes through action recognition, Pan et al. [246] put up a sports referee training system, which could recognize whether a trainee equipped with the Myo armband emits precise judging signals while reviewing a prerecorded professional game. In this study, they employed deep belief networks (DBNs) to extract more representative features for hand gesture recognition.

5.1.2 Video segments highlights

The segmentation and summary of highlights in sports videos has attracted a large audience and possess significant commercial potential. In order to achieve this goal, the fundamental step is to apply action recognition into processing the video. Nakano et al.[247] introduced a brand-new automatic highlight detection method, with the blink rate, identified not just the emblematic athletic movements but also the unique artistic expressions in figure skating performances. By designating these as key frames, the supervised learning approach grounded in blink rate enabled highly accurate

detection that aligns more closely with human perception. Besides, Tian et al. [248] designed a comprehensive, multi-technology framework to acquire a 3D human pose for the analysis of jumps in figure skating, finally serving the purpose of presenting the animated 3D pose of a figure skater.

5.1.3 Training helper

Sports video corpus, with extensive records, is a valuable resource for coaches and players. Video action recognition identifies fundamental sports units. By linking action sequences to winning strategies, coaches can tailor training and devise game plans. Fani et al.[249] proposed a unified architecture as an action recognition hourglass network(ARHN). ARHN has three elements, the first element is the estimator for latent poses, the second converts latent features into a unified frame of reference, and the third component is responsible for recognizing actions. Another popular training helper system is the sports AI coach system, it was presented by Wang et al. [250], which could offer personalized athletic training experiences with multiple unique features, and could demonstrate through comprehensive user studies that the system significantly enhances the user's training experience. Action recognition acts as a crucial step within the AI coach system, enabling the extraction and summary of complex visual information.

5.2 The applications in medical rehabilitation

In recent years, there has been a gradual rise in the number of stroke patients. This has been accompanied by a significant imbalance between the number of rehabilitation physicians and the patients they need to serve, resulting in many patients being unable to undergo the necessary rehabilitation training.

Pan et al. [251] proposed a Kinect camera-based rehabilitation robot schematic action recognition. They adopted Graph Attention neural network to build the skeleton data of the rehabilitation physician. First, motion information is acquired and tracked using Kinect. Next, the behavior of the rehabilitation physician is captured, and skeleton data are generated. The generated data are then processed to attenuate jitter. Following this, the behavior of the demonstration is mapped. The robotic arm is then modeled and forward kinematics are calculated. Finally, the angles are calculated. Besides, Wang et al. [252] established a multi-sensor fusion system(MSFS) for rehabilitation robots. The system possesses the capability to precisely determine the patient's behavioral state and offer a supplementary basis for judging the formulation of rehabilitation strategies employing rehabilitation robots. Additionally, it furnishes the underlying controller with the requisite control input.

5.3 The applications in traffic safety

Traffic safety is an essential issue all over the world. Most road accidents are primarily associated with the driver's dangerous driving behaviors and the improper actions from the pedestrian. In this section, we would introduce the applications of human action recognition in traffic safety with two subsections.

5.3.1 Drivers' behaviors

Yan et al. [253] established a vision-based network to recognize the driver's actions using CNNs. Specifically, when presented with an image, regions resembling skin were extracted using a Gaussian Mixture Model. These extracted regions were then fed into a deep convolutional neural network model, named R*CNN, for the purpose of generating action labels. The skin-like regions offered a wealth of semantic information, possessing significant discriminative power. What's more, R*CNN effectively selects the most informative regions from a pool of candidates, thereby enhancing the final action recognition process. Besides, to make the algorithm more applicable, Zhao et al. [254] put forward a brand-new driver distracting behavior monitoring system through a wearable acoustic sensor that utilizes the Mel-spectrum transform with deep learning techniques. In order to make the recognizing system real-time, Seong et al. [255] proposed a deep learning-based neural architecture search(NAS) to distinguish the actions of the drivers. In the NAS method, a reinforcement learning algorithm was employed to efficiently search for neural network architectures by leveraging shared parameter weights. In addition, they compiled their own dataset for classifying driver behavior, identifying four prevalent driving habits: standard driving, using a mobile phone, eating, and smoking.

5.3.2 Pedestrians' behaviors

The contemporary world is marked by significant advancements in innovative technology and intelligent machinery, along with improvements in transport networks, emergency response systems, and educational facilities. Yet, effectively comprehending complex environments, conducting crowd monitoring, and observing individual behaviors remain formidable challenges, especially the pedestrians on the road in modern cities.

Akhter et al. [257] presented an outstanding architecture for understanding the crowd data via a sustainable framework and for classifying pedestrian anomalous and normal actions. In particular, they used smart graph to optimize the whole network. With the development of economic industry, the issue of pedestrians using smartphones while walking has gained prominence as a hazardous behavior, and grasping the underlying factors of such conduct is vital for minimizing traffic accidents. Ban et al.[256] used a logistic regression model to represent the pedestrian's decision, and the decision's ambiguity was assessed using information entropy derived from this model. In addition, statistical analysis was conducted on pedestrian movement to examine the influence of smartphone use while walking on their decisions and movements. Fig. 16 shows the details of [256]. What's more, Pang et al. [258] introduced a method for trajectory prediction, named BR-GAN, which combines geographical, social, and behavioral context-awareness. During the trajectory decision-making process, three key constraints are introduced: geographical, social, and intention constraints. To account for the uncertainty in predicting future paths, BR-GAN incorporates these constraints into its prediction framework based on GANs. In this prediction process, the model comprehensively considers geographical interactions, social interactions, and intention estimations.

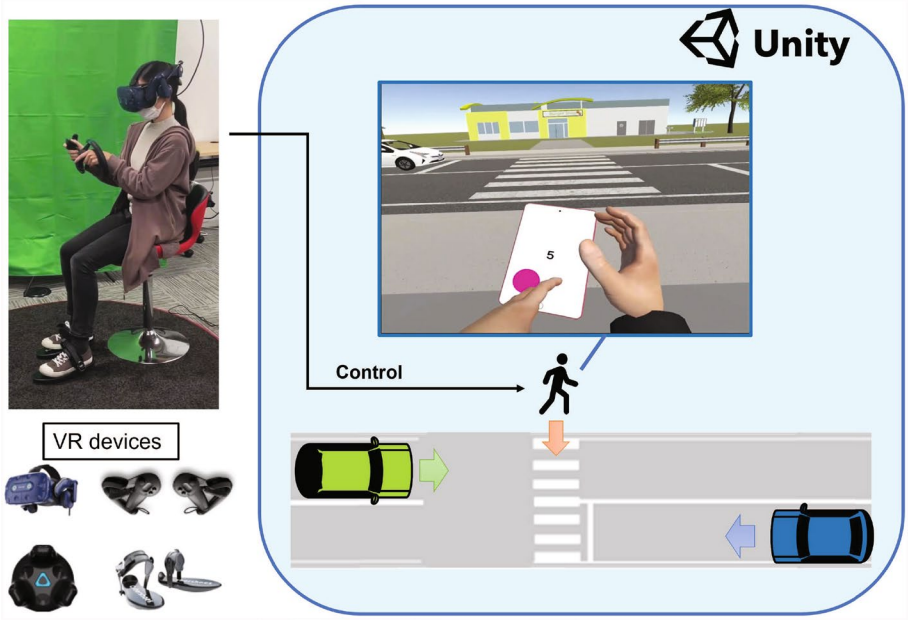

Fig. 16 The details of [256]. Figure from [256]

It further employs an attention mechanism to select feasible and interpretable paths from the potential options (Table 11).

6 Future directions

From the five sections before, we have discussed the recent human action recognition algorithms, the benchmark datasets, the combination with other fields and the applications of human action recognition, this section we would give you some future research directions.

6.1 Multimodal fusion

The pursuit of robust multimodal fusion in HAR confronts two fundamental challenges: asynchronous sensor integration and nonlinear cross-modal interactions.

A typical work for the former one is proposed by Wang et al. [264], which introduces a hierarchical multimodal framework to resolve sensor asynchrony via contrastive semantic alignment. Experiments show that such hierarchical processing—operating at data-level (modality-specific encoding), feature-level (semantic alignment), and decision-level (adaptive fusion)—improves recognition accuracy by 19% on datasets with heterogeneous sampling rates, demonstrating the efficacy of structured fusion pipelines in handling real-world sensor disparities.For the latter one, Liu et al. [263] employs self-supervised factorization to separate sentiment-relevant features from modality-specific noise, aligning latent spaces via adversarial learning to reconstruct missing modalities. Experimental results indicate that this approach boosts robustness to partial data by 25% on challenging benchmarks, directly addressing a common bottleneck in HAR where sensor failures or occlusions often degrade performance. This aligns with the self-supervised co-learning paradigm, where GAN-based or mutual information maximization techniques enhance resilience to incomplete data–an essential capability for real-world

Table 11 Comparison for the methods about the applications of HAR

Basement	Methods	Category	Metric[1](%)	Key contribution
Sports(video judge)	FALCONS [244]	Diving	84.53 UNLV-diving [259]	Adopt ST-GCN and HRNet to assess the actions
	Pan et al. [246]	Basketball	91.01 Myo [246]	Set up a sports referee Deep belief networks (DBNs)
Sports(video segments)	Nakano et al. [247]	Hockey	78 HARPE [247]	ARHN architecture Pose-based feature fusion
	Tian et al. [248]	Figure skating	87.25 Figure skating[248]	Multi-perspective stereo reconstruction
Sports(training helper)	ARHN [249]	Hockey	78 HARPE [247]	ARHN architecture Pose-based feature fusion
	AI Coach [250]	Freestyle skiing aerials	83.7 Freestyle skiing aerials[250]	Action recognition acts as a crucial step within the AI coach system
Medical rehabilitation	Pan et al. [251]	Rehabilitation robotics	0.383(stationary state) 8.864(movement state)[2]	Human–robot motion mapping method
	Wang et al. [252]	Rehabilitation robotics	BP neural network classification accuracy: 92.38	Multi-sensor fusion system Two-level neural network
traffic(Driver)	Yan et al. [253]	Driver action recognition	97.76 SEU [260]	Vision-based network Skin color modeling
	Zhao et al. [254]	Distracted driving actions	Average F1-value: 91.32	Privacy-friendly sensing framework Multi-class distraction recognition Acoustic signal advantages
	NAS [255]	Driver action recognition	91.12(Model 1) 92.08(Model 2)	RL-based network for real-time HAR 4-class driver behavior dataset
traffic(pedestrians)	Akhter et al. [257]	Normal/abnormal action recognition	83.83 UMN [261]	Graph-based optimization framework Graph mining algorithm
	Ban et al. [256]	Smartphone/normal walking	3.17s(Without smartphone) 1.67s(With smartphone)[3]	VR-based simulator for HAR Logistic regression model
	BR-GAN [258]	Pedestrian trajectory prediction	ADE 0.73m FDE 1.37m ETH [262]	Action recognition integration Multi-attention mechanism

[1] Unless otherwise stated, the results are validated on the dataset made by the authors.

[2] The jitter of wrist skeletal points in the stationary state is reduced by approximately 48.2% and the jitter of wrist skeletal points in the movement state is reduced by approximately 5.4%.

[3] Recognition time per trial before crossing.

HAR systems often plagued by sensor dropout or noise. As a result, there still remains opportunities to design more fusion and co-learning strategies for multi-modality human action recognition.

Multimodal fusion in human action recognition (HAR) typically employs hierarchical structured fusion (data-level alignment, feature-level semantic integration, and decision-level adaptive weighting) and self-supervised co-learning (factorized noise separation and adversarial modality alignment), as demonstrated by Wang et al. [264] and Liu et al. [263], who achieved 19% and 25% accuracy improvements, respectively. However, challenges persist: sensor asynchrony (e.g., temporal misalignment between RGB and IMU data) risks action misinterpretation, requiring solutions like dynamic time warping (DTW); modality-specific noise (e.g., low-light infrared artifacts) degrades robustness, necessitating self-supervised denoising or GAN-based reconstruction; computational complexity hinders real-time deployment, demanding lightweight architectures (e.g., MobileNetV3) or hardware acceleration; and interpretability gaps in models like Transformers limit trustworthiness, urging interpretable frameworks (e.g., Grad-CAM). Future work should focus on cross-modal graph neural networks (GATs), edge-optimized models, and rule-constrained fusion strategies to address these limitations in real-world HAR systems.

6.2 Improvement of hardware acceleration and edge computing

The high computational cost of 3D CNNs in human action recognition (HAR) necessitates hardware acceleration and edge computing. By combining hardware acceleration technologies, such as GPU, FPGA (Field Programmable Gate Array), ASIC (Application Specific Integrated Circuit), etc., and edge computing architectures, part of the computing tasks can be transferred from the cloud to edge devices for processing. Fan et al. [265] propose a unified FPGA architecture with static Block Floating Point (BFP) quantization, eliminating frequent FP-BFP conversions to achieve 92.4%/85.2% MAC efficiency for 2D/3D CNNs. This design reduces logic resource usage by 50% compared to integer quantization while maintaining 8-bit mantissa accuracy (e.g., 90.077% accuracy on UCF101 for C3D). By integrating edge-compatible hardware acceleration, the architecture achieves 1667 GOP/s throughput on ResNet-50 with 37.0 GOP/s/W energy efficiency, enabling real-time HAR in IoT, smart security, and industrial applications with low latency (e.g., 93.95 ms for C3D on UCF101). This work bridges the gap between high-performance HAR models and edge deployment feasibility.

The integration of hardware acceleration (e.g., FPGA, ASIC) and edge computing for HAR has focused on optimizing energy efficiency, latency, and scalability through hardware-aware model compression (dynamic quantization, pruning, knowledge distillation), edge-centric architectures (heterogeneous edge-cloud collaboration, neuromorphic computing), and multi-modal fusion strategies (cross-modal attention, privacy-preserving inference). For instance, dynamic BFP quantization in 3D CNNs achieves 92.4% MAC efficiency with 8-bit mantissa accuracy [265], while lightweight models like MobileNetV3 reduce FLOPs by 80% without significant accuracy loss. Challenges such as sensor asynchrony (e.g., RGB-D/IMU mismatch) are addressed via time-aware attention or event-driven processing, and memory constraints are mitigated through compressed inference (e.g., tensor decomposition). Emerging solutions include self-calibrating accelerators (adaptive precision), low-latency pipeline parallelism (e.g., 93.95 ms on UCF101), and edge-optimized multi-modal systems (e.g., MM-GAT on

FPGAs). Future directions prioritize hardware–software co-optimization (domain-specific ASICs), privacy-preserving edge inference, and adaptive architectures to balance real-time performance, energy efficiency, and robustness in resource-constrained applications (e.g., IoT, smart cities).

6.3 Improvement of group interaction action recognition

With the increasing application of RGB-D data in fields such as social scenes and sports events, multi-person interaction behavior recognition has become an important research direction. Mao et al. [266] introduced a multi-scale Sub-group Context Block (SCB) for group action recognition.SCB utilized an assignment matrix to automatically learn the mapping from actors to sub-groups, enabling the system to automatically capture the representation and interactions within each subgroup. SAB employs deep reinforcement learning to score sub-groups dynamically, emphasizing critical ones for improved discriminability.By fusing features from SCBs with varying clustering numbers (e.g., 2/3/4/6 sub-groups), the model captures multi-scale context—from team-level strategies to small-group collaborations. Moreover, Zhu et al. [267] proposed Hierarchical Spatial–Temporal Transformer termed HSTT for group action recognition, focusing on capturing the various degrees of spatial–temporal dynamic interactions adaptively and jointly among actors.Thus, the technical implementation paths in multi-person interaction recognition include multi-scale subgroup modeling (e.g., Mao et al.'s Sub-group Context Block (SCB) with dynamic subgroup scoring via deep reinforcement learning (SAB) [266]), which captures hierarchical interactions from team-level strategies to small-group collaborations by fusing features across varying clustering numbers (2/3/4/6 sub-groups). Additionally, hierarchical spatial–temporal Transformers (e.g., Zhu et al.'s HSTT [267]) adaptively model diverse spatial–temporal interactions through stacked spatial graph and local–global temporal transformer blocks. Potential challenges include sensor synchronization issues (e.g., RGB-LiDAR data alignment), scalability for large groups (e.g., computational complexity in clustered sub-group analysis), robustness in dynamic scenes (e.g., occlusion handling and motion trajectory extraction), and real-time deployment constraints due to high computational demands of attention mechanisms. Cross-disciplinary innovations in sensor fusion, lightweight architectures, and temporal reasoning are critical to address these challenges. As the demand for intelligent analysis in social scenarios and sports events grows, developing models and algorithms for multi-person behavior recognition has emerged as an urgent frontier within the broader field of human action recognition, requiring simultaneous handling of collective behaviors, discrimination of individual actions, and modeling of interactive relationships. This entails multidimensional challenges such as robust human tracking across dynamic scenes, fine-grained action separation in occluded environments, and spatio-temporal interaction modeling that captures hierarchical group dynamics—challenges that necessitate cross-disciplinary innovations in computer vision, graph neural networks, and temporal reasoning.

6.4 Construction of large-scale and complex datasets

Currently, the existing RGB-D HAR datasets still face critical limitations, such as insufficiently diverse scenes, limited action categories, and relatively small data scales,

e.g., N-UCLA [142], hindering . To address this, constructing larger-scale datasets with diverse scenes (indoor/outdoor, crowded), rich interaction patterns (cooperative/competitive behaviors), and multi-modal annotations (3D joints, social relations) is essential. Improved annotation quality (e.g., inter-annotator agreement ≥ 0.85) and standardized formats will enhance supervision accuracy, enabling advanced models like GCNs to capture hierarchical dynamics and drive applications in smart surveillance and sports analytics.

The construction of large-scale, complex HAR datasets involves multi-modal data acquisition (e.g., RGB-D, LiDAR, IMU) to capture spatio-temporal dynamics and 3D structure, diverse scene design (indoor/outdoor, crowded/non-crowded) with rich interaction patterns (cooperative/competitive/ambiguous actions), and high-quality annotations (e.g., 3D joints, social relations) via semi-automated tools (e.g., OpenPose) and standardized formats. Scalable data augmentation (e.g., GANs for synthetic motion trajectories) ensures dataset diversity while preserving semantic consistency. However, challenges persist in sensor synchronization (e.g., RGB-LiDAR misalignment), annotation consistency for subjective actions, computational demands for large-scale data processing, cross-modal feature alignment, and privacy constraints in real-world deployment. Addressing these requires innovations in sensor fusion, distributed frameworks, and privacy-preserving techniques to enable robust model training and real-world applications.

7 Conclusion

HAR has emerged as a cornerstone of computer vision, with significant advancements in multimodal algorithms, benchmark datasets, and cross-domain applications.In this survey, we comprehensively review the RGB-D-based methods for video action recognition, including the SOTA algorithms in different modalities, the benchmark datasets and the SOTA methods for each dataset, the combined algorithms in different subfields, the applications of HAR in various areas and the possible promising research directions for HAR.

This survey highlights the evolution of HAR revolutions among four modalities, RGB, depth, skeleton and RGB-D. Modern algorithms, such as 3D CNNS, two-stream networks and attention-based models, have demonstrated remarkable performance in capturing spatial–temporal features. In this survey, we give a detailed introduction for the combination between HAR and object detection and object tracking, which serves as a fundamental solution for the task of computer vision.Additionally, the applications of HAR in sports(video judge, video segments highlights and training helper), medical rehabilitation(rehabilitation robots) and traffic safety(drivers' behaviors and pedestrians' behaviors) proves the significance and versatility of HAR in recent research of computer vision. Moreover, for the future research directions, we summarize three core ones, multi-modal fusion, improvement of hardware acceleration and edge computing, improvement of group interaction action recognition and construction of novel large-scale HAR dataset.

Abbreviations

CNN Convolutional neural networks
RNN Recurrent neural networks

GNN Graph neural networks
GCN Graph convolutional networks

Acknowledgements
No additional acknowledgements.

Author contributions
All authors participated in the research of the recent works for human action recognition and writing of the manuscript. All authors read and approved the final manuscript.

Funding
Not applicable.

Data availability
This is a review article, and no new data were generated or analyzed. Therefore, this section is not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 16 January 2025 Accepted: 10 July 2025
Published online: 13 August 2025

References

1. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in *2014 IEEE conference on computer vision and pattern recognition*. ed. by A. Karpathy (IEEE, Columbus, 2014), pp.1725–1732
2. K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in *Advances in neural information processing systems*, vol. 27, ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K.Q. Weinberger (Curran Associates Inc, New york, 2014)
3. L. Wang, Y. Qiao, X. Tang, Action recognition with trajectory-pooled deep-convolutional descriptors. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), (2015), pp.4305–4314
4. J. Donahue, L.A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 677–691 (2017)
5. A. Barkoky, N.M. Charkari, Complex network-based features extraction in RGB-D human action recognition. J. Vis. Commun. Image Represent. **82**, 103371 (2022)
6. P. Khaire, P. Kumar, Deep learning and rgb-d based human action, human-human and human-object interaction recognition: a survey. J. Vis. Commun. Image Represent. **86**, 103531 (2022)
7. Z. Liu, J. Cheng, L. Liu, Z. Ren, Q. Zhang, C. Song, Dual-stream cross-modality fusion transformer for RGB-D action recognition. Knowl. Based Syst. **255**, 109741 (2022)
8. D. Liu, F. Meng, Q. Xia, Z. Ma, J. Mi, Y. Gan, M. Ye, J. Zhang, Temporal cues enhanced multimodal learning for action recognition in RGB-D videos. Neurocomputing **594**, 127882 (2024)
9. L. Chen, H. Wei, J. Ferryman, A survey of human motion analysis using depth imagery. Pattern Recogn. Lett. **34**(15), 1995–2006 (2013)
10. J.K. Aggarwal, L. Xia, Human activity recognition from 3d data: a review. Pattern Recogn. Lett. **48**, 70–80 (2014)
11. H. Cheng, L. Yang, Z. Liu, Survey on 3d hand gesture recognition. IEEE Trans. Circuits Syst. Video Technol. **26**(9), 1659–1673 (2016)
12. S. Escalera, V. Athitsos, I. Guyon, Challenges in multi-modal gesture recognition. J. Mach. Learn. Res. **1**, 1–60 (2017)
13. J. Zhang, W. Li, P.O. Ogunbona et al., Rgb-d-based action recognition datasets: a survey. Pattern Recogn. **60**, 86–105 (2016)
14. F. Han, B. Reily, W. Hoff et al., Space-time representation of people based on 3d skeletal data: a review. Comput. Vis. Image Underst. **158**, 85–105 (2017)
15. G. Yao, T. Lei, J. Zhong, A review of convolutional-neural-network-based action recognition. Pattern Recogn. Lett. **118**, 14–22 (2019)
16. Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, J. Liu, Human action recognition from various data modalities: a review. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 3200–3225 (2023)
17. B. Liu, H. Cai, Z. Ju et al., RGB-D sensing based human action and interaction analysis: a survey. Pattern Recogn. **94**, 1–12 (2019)
18. R. Singh, A. Sonawane, R. Srivastava, Recent evolution of modern datasets for human activity recognition: a deep survey. Multimed. Syst. **26**, 83–106 (2019)
19. L.L. Presti, M.L. Cascia, 3d skeleton-based human action classification: a survey. Pattern Recogn. **53**, 130–147 (2016)

20. J. Sedmidubsky, P. Elias, P. Budikova, P. Zezula, Content-based management of human motion data: survey and challenges. IEEE Access **9**, 64241–64255 (2021)
21. O. Oreifej, Z. Liu, Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences, in: 2013 IEEE conference on computer vision and pattern recognition, (2013), pp. 716–723
22. C. Chen, K. Liu, N. Kehtarnavaz, Real-time human action recognition based on depth motion maps. J. Real-Time Image Proc. **12**, 155–163 (2016)
23. X. Yang, Y. Tian, Effective 3d action recognition using eigenjoints. Vis. Commun Image Represent. **25**(1), 2–11 (2014)
24. M. Li, H. Leung, H.P. Shum, Human action recognition via skeletal and depth based feature fusion, in *Proceedings of the 9th international conference on motion in games*. ed. by M. Li (ACM Digital Library, New York, 2016)
25. H. Wang, A. Kläser, C. Schmid, C.-L. Liu, Action recognition by dense trajectories. CVPR **1**, 3169–3176 (2011)
26. H. Wang, C. Schmid, Action recognition with improved trajectories, in *2013 IEEE International conference on computer vision*. ed. by H. Wang (Open access, New york, 2013), pp.3551–3558
27. X. Peng, C. Zou, Y. Qiao et al., Action recognition with stacked fisher vectors, in *Proceedings of the European conference on computer vision*. ed. by X. Peng (Springer, Cham, 2014)
28. Z. Lan, M. Lin, X. Li et al., Beyond gaussian pyramid: multi-skip feature stacking for action recognition, in *2015 IEEE conference on computer vision and pattern recognition (CVPR)*. ed. by Z. Lan (CV Foundation, New Delhi, 2015)
29. H. Bay, T. Tuytelaars, L.V. Gool, SURF: Speeded up robust features, in *Proceedings of the European conference on computer vision*. ed. by H. Bay (Springer, Berlin, 2006)
30. K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos. Adv. Neural. Inf. Process. Syst. **27**, 1 (2014)
31. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. Commun. ACM **60**, 84–90 (2012)
32. L. Wang, Z. Wang, Y. Xiong et al., CUHK and SIAT Submission for THUMOS 15 Action Recognition Challenge, in *Proceedings of the THUMOS 15 action recognition challenge* (2015)
33. K. Soomro, A.R. Zamir, M. Shah, Ucf101: A dataset of 101 human actions classes from videos in the wild. Comput. Sci. (2012). https://arxiv.org/abs/1212.0402
34. L. Wang, Y. Xiong, Z. Wang, Y. Qiao, Towards good practices for very deep two-stream convnets. Comput. Sci. (2015). https://arxiv.org/abs/1507.02159
35. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint. (2014) arXiv:1409.1556
36. C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in *IEEE Conference on computer vision and pattern recognition (CVPR)*. ed. by C. Feichtenhofer (IEEE, New york, 2016), pp.1933–1941
37. M. Yang, Y. Guo, F. Zhou, Z. Yang, Ts-d3d: A novel two-stream model for action recognition, in *2022 International conference on image processing, computer vision and machine learning (ICICML)*. ed. by M. Yang (IEEE, Xi'an, 2022), pp.179–182
38. G. Chéron, I. Laptev, C. Schmid, *P-cnn: Pose-based CNN features for action recognition*. In: IEEE International conference on computer vision (2015)
39. L. Wang, Y. Xiong, Z. Wang et al., Temporal segment networks: towards good practices for deep action recognition
40. M. Umran, K. Muchtar, T.F. Abidin, F. Arnia, Action localization and recognition through unsupervised i3d and TSN, In: 2023 3rd International conference on computing and information technology (ICCIT), (2023), pp. 269–273
41. J. Carreira, A. Zisserman, Q. Vadis, Action recognition? a new model and the kinetics dataset, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017)
42. S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)
43. D. Tran, L. Bourdev, R. Fergus, et al., Learning spatiotemporal features with 3d convolutional networks, in *Proceedings of the IEEE international conference on computer vision* (2015)
44. L. Yao, A. Torabi, K. Cho et al. Describing videos by exploiting temporal structure, in *2015 IEEE International Conference on Computer Vision* (ICCV), (2015) pp. 4507–4515
45. J. Deng, W. Dong, R. Socher et al. Imagenet: A large-scale hierarchical image database, in *2009 IEEE conference on computer vision and pattern recognition*, (2009), pp. 248–255
46. B. Zhou, A. Lapedriza, A. Khosla et al., Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 1452–1464 (2018)
47. K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?, in *2018 IEEE/CVF conference on computer vision and pattern recognition*, (2017), pp. 6546–6555
48. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE conference on computer vision and pattern recognition (CVPR)*. (2016). pp. 770–778
49. S. Xie, R.B. Girshick, P. Dollár, et al. Aggregated residual transformations for deep neural networks. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (2016), pp. 5987–5995
50. J. Hu, L. Shen, S. Albanie, et al., Squeeze-and-excitation networks, in *2018 IEEE/CVF conference on computer vision and pattern recognition*. (2017). p. 7132–7141
51. Diba A, Fayyaz M, Sharma V, et al. Spatio-temporal channel correlation networks for action classification, in *Proceedings of the European Conference on Computer Vision* (2018)
52. Z. Qiu, T. Yao, T. Mei, Learning spatio-temporal representation with pseudo-3d residual networks, in 2017 IEEE International Conference on Computer Vision (ICCV), (2017), pp. 5534–5542
53. D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A closer look at spatiotemporal convolutions for action recognition, pp. 6450–6459 (2018)
54. Y. Zhao, Y. Xiong, D. Lin, Trajectory convolution for action recognition. Adv. Neural. Inf. Process. Syst. **31**, 1 (2018)
55. Y. Zhou, X. Sun, Z.-J. Zha, W. Zeng, Mict: Mixed 3d/2d convolutional tube for human action recognition, in: *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 449–458 (2018)

56. L. Wang, W. Li, W. Li, L. Van Gool: Appearance-and-relation networks for video classification, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2018), pp. 1430–1439
57. S. Xie, C. Sun, J. Huang, et al. Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification, in *Proceedings of the European conference on computer vision* (2017)
58. Y. Men, J. Luo, Z. Zhao, H. Wu, F. Luo, G. Zhang, M. Yu, Surgical gesture recognition in open surgery based on 3DCNN and slowfast, in *2024 IEEE 7th information technology, networking, electronic and automation control conference (ITNEC)*, vol. 7, ed. by Y. Men (IEEE, Chongqing, 2024), pp.429–433
59. H.-H. Chang, Y.-H. Chang, Y.-L. Shih, C.-H. Lin, H.-C. Shih, Basketball player action recognition and tracking using r(2+1)d CNN with spatial-temporal features, in *2024 IEEE 13th global conference on consumer electronics (GCCE)*. ed. by H.H. Chang (IEEE, Kitakyushu, 2024), pp.388–389
60. N. Murugan, S. Sathasivam, Real-time human action recognition by using r(2+1)d convolutional neural network. In: 2024 3rd international conference on artificial intelligence for internet of things (AIIoT), (2024) pp.1–6
61. G. Varol, I. Laptev, C. Schmid, Long-term temporal convolutions for action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 1510–1517 (2016)
62. A. Diba, M. Fayyaz, V. Sharma, et al., Temporal 3D ConvNets: new architecture and transfer learning for video classification. ArXiv. https://arxiv.org/abs/1711.08200 (2017)
63. G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), (2017), pp. 2261–2269
64. X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in *2018 IEEE/CVF conference on computer vision and pattern recognition*, (2018), pp.7794–7803
65. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł Kaiser, I. Polosukhin, Attention is all you need. Adv. Neural. Inf. Process. Syst. **30**, 1 (2017)
66. L. Zhu, D. Tran, L. Sevilla-Lara, et al., Faster recurrent networks for efficient video classification, in *Proceedings of the AAAI conference on artificial intelligence* (2020)
67. M.-H. Ha, Top-heavy capsnets based on spatiotemporal non-local for action recognition. J. Comput. Theor. Appl. **2**(1), 39–50 (2024)
68. N. Elmadany, L. Gao, L. Guan, A non local multi-fiber network for action anticipation in videos, in *2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, (2022) pp. 384–388
69. Z. Dong, Fast action recognition based on local and nonlocal temporal feature, in *2021 IEEE 4th International conference on information systems and computer aided education (ICISCAE)*, (2021), pp. 196–200
70. O. KöPüKLü, N. Kose, A. Gunduz, et al., Resource efficient 3d convolutional neural networks, in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, (2019), pp.1910–1919
71. D. Tran, H. Wang, M. Feiszli, L. Torresani, Video classification with channel-separated convolutional networks, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), pp.5551–5560
72. Y. Chen, Y. Kalantidis, J. Li, S. Yan, J. Feng, Multi-fiber networks for video recognition, in, Proceedings of the European conference on computer vision (ECCV), (2018), pp.352–367
73. C. Feichtenhofer, H. Fan, J. Malik, et al. Slowfast networks for video recognition, in *2019 IEEE/CVF international conference on computer vision (ICCV)*, (2018), pp.6201–6210
74. Q. Fan, C.-F.R. Chen, H. Kuehne, M. Pistoia, D. Cox, More is less: Learning efficient video representations by big-little network and depthwise temporal aggregation. Adv. Neural. Inf. Process. Syst. **32**, 1 (2019)
75. T. Gopalakrishnan, N. Wason, R.J. Krishna, N. Krishnaraj, Comparative analysis of fine-tuning i3d and slowfast networks for action recognition in surveillance videos. Eng. Proc. **59**(1), 203 (2024)
76. J. Donahue, L.A. Hendricks, M. Rohrbach, et al. Long-term recurrent convolutional networks for visual recognition and description, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2014), pp.2625–2634
77. J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, et al. Beyond short snippets: deep networks for video classification, in *2015 IEEE conference on computer vision and pattern recognition (CVPR)*, (2015), pp.4694–4702
78. Q. Li, Z. Qiu, T. Yao, et al. Action recognition by learning deep multi-granular spatio-temporal video representation, in *Proceedings of the 2016 ACM on international conference on multimedia retrieval* (2016)
79. Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, C.G. Snoek, Videolstm convolves, attends and flows for action recognition. Comput. Vis. Image Underst. **166**, 41–50 (2018)
80. L. Sun, K. Jia, K. Chen, D.Y. Yeung, B.E. Shi, S. Savarese, Lattice long short-term memory for human action recognition, in: *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), pp. 166–2175
81. Y. Shi, Y. Tian, Y. Wang, et al. Learning long-term dependencies for action recognition with a biologically-inspired deep network, in *2017 IEEE International conference on computer vision (ICCV)*, (2016), pp.716–725
82. SS. Uday, ST. Pavani, TJ. Lakshmi, R. Chivukula, Classifying human activities using machine learning and deep learning techniques. arXiv preprint arXiv:2205.10325 (2022)
83. H. Gammulle, S. Denman, S. Sridharan, et al., Two stream lstm: a deep fusion framework for human action recognition, in *2017 IEEE winter conference on applications of computer vision (WACV)*, (2017), pp.177–186
84. I. Sheth, Three-stream network for enriched action recognition. arXiv preprint arXiv:2104.13051 (2021)
85. D. Tilley, U. Martinez-Hernandez, Shallow hierarchical CNN-LSTM for activity recognition to integrate postural transition states. In: 2023 IEEE Sensors, (2023), pp. 1–4
86. Z. Yin, C. Li, X. Dong, Video rwkv: Video action recognition based RWKV. arXiv preprint arXiv:2411.05636 (2024)
87. L. Zhu, D. Tran, L. Sevilla-Lara, Y. Yang, M. Feiszli, H. Wang, Faster recurrent networks for efficient video classification, In *Proceedings of the AAAI conference on artificial intelligence*, **34**, pp. 13098–13105 (2020)
88. G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding? ICML **2**, 4 (2021)
89. D. Neimark, O. Bar, M. Zohar, et al. Video transformer network, in *2021 IEEE/CVF international conference on computer vision workshops* (ICCVW), (2021), pp.3156–3165
90. Y. Cao, J. Xu, S. Lin, et al. Gcnet: Non-local networks meet squeeze-excitation networks and beyond, in *Proceedings of the IEEE/CVF international conference on computer vision workshops* (2019)
91. E. Khazaei, A. Esmaeilzehi, B. Taha, D. Hatzinakos, Cdfl: Efficient federated human activity recognition using contrastive learning and deep clustering. IEEE Sens. J. **24**(22), 38196–38208 (2024)

92. Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video swin transformer, in: *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (2022), pp.3192–3201
93. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), pp.10012–10022
94. K. Doshi, Y. Yilmaz, Semantic video transformer for robust action recognition, in *2023 IEEE Conference on Dependable and Secure Computing (DSC)*, (2023), pp.1–5
95. Y. Jing, F. Wang, Tp-vit: a two-pathway vision transformer for video action recognition, in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2022), pp.2185–2189
96. Y. Ren, C. Li, W. Bao, X. Chen, Y. Jing, A study of student action recognition in smart classrooms based on improved slowfast swin transformer, in *2023 8th International conference on signal and image processing (ICSIP)*, (2023), pp.59–63
97. C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in *Proceedings of the 17th international conference on pattern recognition* (2004)
98. K.K. Reddy, M. Shah, Recognizing 50 human action categories of web videos. Mach. Vis. Appl. **24**(5), 971–981 (2013)
99. J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos "in the wild", in *2009 IEEE conference on computer vision and pattern recognition*, (2009), pp. 1996–2003
100. F.C. Heilbron, V. Escorcia, B. Ghanem, et al. Activitynet: a large-scale video benchmark for human activity understanding, in *2015 IEEE conference on computer vision and pattern recognition (CVPR)*, (2015), pp.961–970
101. H. Jhuang, J. Gall, S. Zuffi, C. Schmid, M.J. Black, Towards understanding action recognition, in *2013 IEEE International Conference on Computer Vision*, (2013), pp.3192–3199
102. W. Kay, J. Carreira, K. Simonyan, et al. The kinetics human action video dataset. ArXiv:1705.06950 (2017)
103. J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, A. Zisserman, A short note about kinetics-600. arXiv preprint arXiv:1808.01340 (2018)
104. J. Reyes-Ortiz, D. Anguita, A. Ghio, L. Oneto, X. Parra, Human activity recognition using smartphones. Esann **3**, 3 (2013)
105. J. Materzynska, G. Berger, I. Bax, R. Memisevic, The jester dataset: a large-scale video dataset of human gestures, in *Proceedings of the IEEE/CVF international conference on computer vision workshops*,(2019)
106. B. Yao, L. Fei-Fei, Grouplet: a structured image representation for recognizing human and object interactions, in *2010 IEEE computer society conference on computer vision and pattern recognition*, (2010), pp.9–16
107. D. Shao, Y. Zhao, B. Dai, D. Lin, Finegym: a hierarchical video dataset for fine-grained action understanding, in *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (2020), pp.2613–2622
108. X. Yang, Y. Tian, Super normal vector for activity recognition using depth sequences. in *2014 IEEE conference on computer vision and pattern recognition*, (2014), pp.804–811
109. H. Liu, L. Tian, M. Liu, et al. SDM-BSM: a fusing depth scheme for human action recognition. *2015 IEEE International conference on image processing (ICIP)* (2015)
110. H. Rahmani, A. Mahmood, D.Q. Huynh, A. Mian, Real time action recognition using histograms of depth gradients and random decision forests, in *IEEE winter conference on applications of computer vision*, (2014), pp.626–633
111. X. Ji, J. Cheng, W. Feng, Spatio-temporal cuboid pyramid for action recognition using depth motion sequences, in *2016 eighth international conference on advanced computational intelligence (ICACI)* (2016)
112. X. Yang, C. Zhang, Y. Tian, Recognizing actions using depth motion maps-based histograms of oriented gradients, in *Proceedings of the 20th ACM international conference on multimedia* (2012)
113. P. Wang, W. Li, Z. Gao et al., Action recognition from depth maps using deep convolutional neural networks. IEEE Trans. Hum. Mach. Syst. **46**(4), 498–509 (2016)
114. Y. Hou, S. Wang, P. Wanga et al., Spatially and temporally structured global to local aggregation of dynamic depth information for action recognition. IEEE Access. **6**, 2206 (2017)
115. H. Rahmani, A.S. Mian, 3d action recognition from novel viewpoints, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), pp.1506–1515
116. K. Sun, B. Xiao, D. Liu, et al. Deep high-resolution representation learning for human pose estimation, in *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (2019), pp.5686–5696
117. J. Shotton, T. Sharp, A. Kipman et al., Real-time human pose recognition in parts from single depth images. CVPR **2011**, 1297–1304 (2011)
118. Y. Du, Y.R. Fu, L. Wang, Skeleton based action recognition with convolutional neural network, in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, (2015), pp.579–583
119. P. Wang, Z. Li, Y. Hou, W. Li, Action recognition based on joint trajectory maps using convolutional neural networks, in *Proceedings of the 24th ACM international conference on multimedia*, (2016), pp.102–106
120. S. Wang, Y. Zhang, H. Qi, M. Zhao, Y. Jiang, Dynamic spatial-temporal hypergraph convolutional network for skeleton-based action recognition, in *2023 IEEE international conference on multimedia and expo (ICME)*, (2023), pp.2147–2152
121. K. Zhu, R. Wang, Q. Zhao et al., A cuboid CNN model with an attention mechanism for skeleton-based action recognition. IEEE Trans. Multimed. **22**(11), 2977–2989 (2020)
122. C. Li, Y. Hou, P. Wang et al., Joint distance maps based action recognition with convolutional neural networks. IEEE Signal Process. Lett. **24**, 624–628 (2017)
123. Q. Ke, S. An et al., Skeletonnet: Mining deep part features for 3-d action recognition. IEEE Signal Process. Lett. **24**, 731–735 (2017)
124. D. Yong, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in *2015 IEEE conference on computer vision and pattern recognition (CVPR)* (2015)
125. G. Pan, Y. Song, S. Wei, Combining pose and trajectory for skeleton based action recognition using two-stream RNN, in *2019 Chinese automation congress (CAC)* (2019)

126. V. Veeriah, N. Zhuang, GJ Qi, Differential recurrent neural networks for action recognition, in *2015 IEEE International conference on computer vision (ICCV)* (2015)
127. P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, N. Zheng, EleATT-RNN: adding attentiveness to neurons in recurrent neural networks. IEEE Trans. Image Process. **29**, 1061–1073 (2020)
128. C. Si, Y. Jing, W. Wang et al., Skeleton-based action recognition with hierarchical spatial reasoning and temporal stack learning network. Pattern Recogn. **107**, 107511 (2020)
129. L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with directed graph neural networks, in *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (2019), pp.904–7913
130. S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in *AAAI conference on artificial intelligence* (2018)
131. F. Ye, H. Tang, Skeleton-based action recognition with JRR-GCN. Electron. Lett. (2019). https://doi.org/10.1049/el.2019.1380
132. C.H. Lin, P.Y. Chou, , C.H. Lin, et al., SlowFast-GCN: a novel skeleton-based action recognition framework, in *2020 International conference on pervasive artificial intelligence (ICPAI)* (2020)
133. W. Yang, J. Zhang, J. Cai et al., Hybridnet: integrating GCN and CNN for skeleton-based action recognition. Appl. Intell. **53**, 574–585 (2022)
134. Y. Zang, D. Yang, T. Liu et al., Sparseshift-GCN: high precision skeleton-based action recognition. Pattern Recogn. Lett. **153**, 136–143 (2021)
135. H. Qiu, Y. Wu, M. Duan, C. Jin, GLTA-GCN: global-local temporal attention graph convolutional network for unsupervised skeleton-based action recognition, in *2022 IEEE international conference on multimedia and expo (ICME)*, (2022), pp.1–6
136. J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in *2012 IEEE conference on computer vision and pattern recognition*, (2012), pp.1290–1297
137. Y.C. Lin, M.C. Hu, W.H. Cheng, Y.H. Hsieh, H.M. Chen, Human action recognition and retrieval using sole depth information, in *Proceedings of the 20th ACM international conference on multimedia*, (2012), pp.1053–1056
138. J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, Robust 3D action recognition with random occupancy patterns, in *Computer Vision – ECCV 2012* (2012)
139. W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3d points, in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, (2010), pp.9–14
140. L. Xia, C.C. Chen, J.K. Aggarwal, View invariant human action recognition using histograms of 3d joints. in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, (2012), pp. 20–27
141. V. Bloom, D. Makris, V. Argyriou, G3d: a gaming action dataset and real time action recognition evaluation framework, in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, (2012), pp.7–12
142. J. Wang, X. Nie, Y. Xia, Y. Wu, S.-C. Zhu, Cross-view action modeling, learning and recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2014), pp.2649–2656
143. J. Kong, Y. Bian, M. Jiang, Mtt: multi-scale temporal transformer for skeleton-based action recognition. IEEE Signal Process. Lett. **29**, 528–532 (2022)
144. Q. Zhang, T. Wang, M. Zhang, K. Liu, P. Shi, H. Snoussi, Spatial-temporal transformer for skeleton-based action recognition, in *2021 China automation congress (CAC)*, (2021), pp.7029–7034
145. A. Shahroudy, J. Liu, T.T. Ng, et al. NTU RGB+D: a large scale dataset for 3D human activity analysis, in *2016 IEEE conference on computer vision and pattern recognition (CVPR)* (2016)
146. L. Xia, C.-C. Chen, J.K. Aggarwal, View invariant human action recognition using histograms of 3d joints, in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, (2012), pp.20–27
147. C. Chen, R. Jafari, N. Kehtarnavaz, UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, in *2015 IEEE international conference on image processing (ICIP)*, (2015), pp.168–172
148. K. Yun, J. Honorio, D. Chattopadhyay, T.L. Berg, D. Samaras, Two-person interaction detection using body-pose features and multiple instance learning, in *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, (2012), pp.28–35
149. G. Zhu, L. Zhang, L. Mei, et al. Large-scale isolated gesture recognition using pyramidal 3D convolutional networks, in *2016 23rd international conference on pattern recognition (ICPR)* (2016)
150. Z. Xu, V. Vilaplana, J.R. Morros, Action tube extraction based 3D-CNN for RGB-D action recognition, in *2018 International conference on content-based multimedia indexing (CBMI)*, (2018), pp.1–6
151. J. Duan, J. Wan, S. Zhou et al., A unified framework for multi-modal isolated gesture recognition. ACM Trans. Multimed. Comput. Commun. Appl. **14**, 1–16 (2018)
152. D. Srihari, P.V.V. Kishore, E.K. Kumar et al., A four-stream convnet based on spatial and depth flow for human action classification using RGB-D data. Multimed. Tools Appl. **79**(17–18), 11723–11746 (2020)
153. P. Wang, W. Li, J. Wan, et al. Cooperative training of deep aggregation networks for RGB-D action recognition, in *AAAI conference on artificial intelligence* (2017)
154. B. Zhou, P. Wang, J. Wan, Y. Liang, F. Wang, D. Zhang, Z. Lei, H. Li, R. Jin, Decoupling and recoupling spatiotemporal representation for RGB-D-based motion recognition, in *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (2022), pp. 20122–20131
155. Z. Shi, T.-K. Kim, Learning and refining of privileged information-based RNNS for action recognition from depth sequences, in *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, (2017), pp.4684–4693
156. B. Mahasseni, S. Todorovic, Regularizing long short term memory with 3d human-skeleton sequences for action recognition, in 2016 *IEEE conference on computer vision and pattern recognition (CVPR)*, (2016), pp.3054–3062
157. X. Xiao, Z. Ren, W. Wei, et al. Shift swin transformer multimodal networks for action recognition in videos, in *2022 International conference on sensing, measurement & data analytics in the era of artificial intelligence (ICSMD)* (2022)
158. D. Wu, L. Pigou, P.J. Kindermans et al., Deep dynamic neural networks for multimodal gesture segmentation and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **38**(8), 1583–1597 (2016)

159. W. Zhu, C. Lan, J. Xing, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks, in *Proceedings of the AAAI conference on artificial intelligence* (2016)
160. Q. Lu, Y. Zhang, M. Lu, V. Roychowdhury, *Action-conditioned on-demand motion generation* (Association for Computing Machinery, New York, 2022)
161. A. Rahnama, A. Mansouri, Temporal relations of informative frames in action recognition, in *2024 13th Iranian/3rd international machine vision and image processing conference (MVIP)* (2024)
162. F. Ronchetti, F.M. Quiroga, L. Lanzarini et al., Distribution of action movements (DAM): a descriptor for human action recognition. Front. Comp. Sci. (2015). https://doi.org/10.1007/s11704-015-4320-x
163. H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, HMDB: a large video database for human motion recognition, in *2011 International conference on computer vision*, (2011), pp.2556–2563
164. L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, Y. Qiao, Videomae v2: scaling video masked autoencoders with dual masking, in *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (2023), pp.14549–14560
165. H. Lu, H. Jian, R. Poppe, et al. Enhancing video transformers for action understanding with VLM-aided training. ArXiv. https://arxiv.org/abs/2403.16128 (2024)
166. B. Sravyapranati, D. Suma, C. Manjulatha et al., Large-scale video classification with convolutional neural networks. Inform. Commun. Technol. Intell. Syst. (2020). https://doi.org/10.1007/978-981-15-7062-9_69
167. D. Tran, H. Wang, M. Feiszli, L. Torresani, Video classification with channel-separated convolutional networks (2019)
168. S. Escalera, X. Baró, J. Gonzalez, M.A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H.J. Escalante, J. Shotton, I. Guyon, Chalearn looking at people challenge 2014: dataset and results, in *Computer vision-ECCV 2014 workshops*, vol. Part 13, ed. by S. Escalera (Springer, Zurich, 2014), pp.459–473
169. P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, J. Kautz, Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks (2016)
170. N. Siddiqui, P. Tirupattur, M. Shah, Dvanet: Disentangling view and action features for multi-view action recognition. Proc. AAAI Conf. Artif. Intell. **38**, 4873–4881 (2024)
171. W. Wu, Z. Sun, W. Ouyang, Revisiting classifier: transferring vision-language models for video recognition. AAAI Conf. Artif. Intell. (2022). https://doi.org/10.1609/aaai.v37i3.25386
172. G.A. Sigurdsson, G. Varol, X. Wang et al., Hollywood in homes: crowdsourcing data collection for activity understanding. Eur. Conf. Comput. Vis. (2016). https://doi.org/10.1007/978-3-319-46448-0_31
173. A. Mondai, S. Nag, J.M. Prada, et al., Actor-agnostic multi-label action recognition with multi-modal query, in *2023 IEEE/CVF International conference on computer vision workshops (ICCVW)* (2023)
174. Q. Cheng, J. Cheng, Z. Liu, Z. Ren, J. Liu, A dense-sparse complementary network for human action recognition based on RGB and skeleton modalities. Exp. Syst. Appl. **244**, 123061 (2024)
175. C. Liu, Y. Hu, Y. Li, et al. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. ArXiv:1703.07475 (2017)
176. Y. Tang, Y. Tian, J. Lu, et al. Action recognition in RGB-D egocentric videos, in *2017 IEEE International conference on image processing (ICIP)*, (2017), pp.3410–3414
177. B. Zhou, P. Wang, J. Wan, Y. Liang, F. Wang, A unified multimodal de- and re-coupling framework for RGB-D motion recognition. IEEE Trans. Pattern Anal. Mach. Intell. **45**(10), 11428–11442 (2023)
178. L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li, C. Liu, M. Liu, Z. Liu, Y. Lu, Y. Shi, L. Wang, J. Wang, B. Xiao, Z. Xiao, J. Yang, M. Zeng, L. Zhou, P. Zhang, Florence: a new foundation model for computer vision, arXiv preprint arXiv:2111.11432 (2021)
179. R. Goyal, S.E. Kahou, V. Michalski, et al. The "Something-something" video database for learning and evaluating visual common sense, in *2017 IEEE international conference on computer vision (ICCV)* (2017)
180. Y. Wang, K. Li, Y. Li, et al. InternVideo: general video foundation models via generative and discriminative learning. ArXiv. https://arxiv.org/abs/2212.03191 (2022)
181. J. Liu, A. Shahroudy, M. Perez et al., NTU RGB+d 120: A large-scale benchmark for 3D human activity understanding. IEEE Trans. Pattern Anal. Mach. Intell. **42**, 2684–2701 (2019)
182. J. Jang, D. Kim, C. Park, et al. Etri-activity3D: a large-scale RGB-D dataset for robots to recognize daily activities of the elderly, in *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, (2020), pp.10990–10997
183. T. Li, J. Liu, W. Zhang, et al. Uav-human: a large benchmark for human behavior understanding with unmanned aerial vehicles, in *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (2021), pp.16261–16270
184. R. Xian, X. Wang, D. Kothandaraman, et al., PMI sampler: patch similarity guided frame selection for aerial action recognition, in *2024 IEEE/CVF winter conference on applications of computer vision (WACV)* (2024)
185. Z. Zhao, Z. Cai, T. Chen, X. Li, H. Li, G. Zhu, Knn-mmd: Cross domain wi-fi sensing based on local distribution alignment. arXiv preprint arXiv:2412.04783 (2024)
186. Z. Zhao, F. Meng, H. Li, X. Li, G. Zhu, Mining limited data sufficiently: a bert-inspired approach for csi time series application in wireless communication and sensing. ArXiv:2412.06861 (2024)
187. L. Keselman, J.I. Woodfill, A. Grunnet-Jepsen, et al. Intel(R) RealSense(TM) stereoscopic depth cameras, in *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (2017)
188. M. Samir, E. Golkar, A.A.A. Rahni, Comparison between the Kinect™ V1 and Kinect™ V2 for respiratory motion tracking, in *2015 IEEE international conference on signal and image processing applications (ICSIPA)* (2015)
189. T. DUTTA, Evaluation of the kinect™ sensor for 3-d kinematic measurement in the workplace. Appl Ergon **43**(4), 645–649 (2012)
190. A.D.C.A. Coroiu, A. Coroiu, Interchangeability of kinect and orbbec sensors for gesture recognition. In: 2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 309–315 (2018). IEEE
191. J. Wan, S.Z. Li, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition, in *2016 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp.761–769

192. B. Ni, G. Wang, P. Moulin, RGBD-HuDaAct: a color-depth video database for human daily activity recognition, in *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, (2011), pp.1147–1153
193. J. Basavaiah, C.G. Patil, Human activity detection and action recognition in videos using convolutional neural networks. J. Inform. Commun. Technol. (2020). https://doi.org/10.32890/jict2020.19.2.1
194. S. Escalera, X. Baró, J. Gonzalez, M.A. Bautista, M. Madadi, M. Reyes, V. Ponce-López, H.J. Escalante, J. Shotton, I. Guyon, Chalearn looking at people challenge 2014: dataset and results, in *Computer vision-ECCV 2014 workshops*, vol. 13, ed. by S. Escalera (Springer, Zurich, 2015), pp.459–473
195. L. Shao, L. Liu, M. Yu, Kernelized multiview projection for robust action recognition. Int. J. Comput. Vis. **118**, 115–129 (2016)
196. J. Basavaiah, C. Patil, Human activity detection and action recognition in videos using convolutional neural networks. J. Inform. Commun. Technol. **19**, 157–183 (2020)
197. V. Veeriah, N. Zhuang, G.-J. Qi, Differential recurrent neural networks for action recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, (2015), pp. 4041–4049
198. M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, A. Baskurt, Sequential deep learning for human action recognition, in *Human behavior understanding: second international workshop, HBU 2011*. ed. by M. Baccouche (Springer, Amsterdam, 2011), pp.29–39
199. M.E. Kalfaoglu, S. Kalkan, A.A. Alatan, Late temporal modeling in 3d CNN architectures with BERT for action recognition, in *Computer Vision – ECCV 2020 Workshops*, (2020), pp. 731–747
200. L. Wang, K. Sun, P. Koniusz, High-order tensor pooling with attention for action recognition. *ICASSP 2024 - 2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (2024), pp. 3885–3889
201. P. Koniusz, L. Wang, A. Cherian, Tensor representations for action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **44**(2), 648–665 (2022)
202. L. Wang, P. Koniusz, Self-supervising action recognition by statistical moment and subspace descriptors, in *Proceedings of the 29th ACM international conference on multimedia*. ed. by L. Wang (Association for Computing Machinery, New York, 2021), pp.4324–4333
203. H. Duan, Y. Zhao, Y. Xiong, W. Liu, D. Lin, Omni-sourced webly-supervised learning for video recognition, in *Computer vision—ECCV 2020: 16th European conference*. ed. by H. Duan (Springer, Glasgow, 2020), pp.670–688
204. S.N. Gowda, M. Rohrbach, L. Sevilla-Lara, Smart frame selection for action recognition. Proc. AAAI Conf. Artif. Intell. (2020). https://doi.org/10.1609/aaai.v35i2.16235
205. W. Wu, X. Wang, H. Luo, J. Wang, Y. Yang, W. Ouyang, Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models, in *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (2023), pp. 6620–6630
206. S. Srivastava, G. Sharma, Omnivec2—a novel transformer based network for large scale multimodal and multitask learning, in *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (2024), pp. 27402–27414
207. Y.H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, G. Toderici, *Beyond short snippets: deep networks for video classification* (IEEE, New York, 2015)
208. D. Wang, W. Ouyang, W. Li, D. Xu, Dividing and aggregating network for multi-view action recognition: 15th European conference, Munich, Germany, September 8–14, 2018, pp.457–473 (Proceedings, Part IX, 2018)
209. F. Baradel, C. Wolf, J. Mille, G.W. Taylor, Glimpse clouds: human activity recognition from unstructured feature points, in *2018 IEEE/CVF conference on computer vision and pattern recognition*, (2018), pp. 469–478
210. S. Vyas, Y.S. Rawat, M. Shah, *Multi-view action recognition using cross-view video prediction* (Springer, Berlin, 2020), pp.427–444
211. S. Das, M.S. Ryoo, Viewclr: Learning self-supervised video representation for unseen viewpoints, in *2023 IEEE/CVF winter conference on applications of computer vision (WACV)*, (2023), pp. 5562–5572
212. K. Shah, A. Shah, C.P. Lau, C.M. Melo, R. Chellapp, Multi-view action recognition using contrastive learning. in *2023 IEEE/CVF winter conference on applications of computer vision (WACV)*, (2023), pp. 3370–3380
213. W. Wu, Y. Zhao, Y. Xu, X. Tan, D. He, Z. Zou, J. Ye, Y. Li, M. Yao, Z. Dong, Y. Shi, Dsanet: Dynamic segment aggregation network for video-level representation learning, in *Proceedings of the 29th ACM international conference on multimedia*, (2021), pp. 1903–1911
214. B. Xia, Z. Wang, W. Wu, H. Wang, J. Han, Temporal saliency query network for efficient video recognition, in *European conference on computer vision*. ed. by B. Xia (Springer, Cham, 2022), pp.741–759
215. B. Xia, W. Wu, H. Wang, R. Su, D. He, H. Yang, X. Fan, W. Ouyang, Nsnet: Non-saliency suppression sampler for efficient video recognition, in *European Conference on Computer Vision*. ed. by B. Xia (Springer, Cham, 2022), pp.705–723
216. Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, Z. Wang, Y. Shi, et al. Internvideo2: scaling foundation models for multimodal video understanding, in *European conference on computer vision*, (2024), pp.396–416
217. B.X.B. Yu, Y. Liu, X. Zhang, S.-H. Zhong, K.C.C. Chan, Mmnet: A model-based multimodal network for human action recognition in RGB-D videos. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 3522–3538 (2023)
218. A. Abdelkawy, A. Ali, A. Farag, Epam-net: an efficient pose-driven attention-guided multimodal network for video action recognition. Neurocomputing **633**, 129781 (2025)
219. D. Reilly, S. Das, Just add π! pose induced video transformers for understanding activities of daily living. In: *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (2024), pp. 18340–18350
220. H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting skeleton-based action recognition. In: 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 2959–2968 (2022). https://doi.org/10.1109/CVPR52688.2022.00298
221. W. Wu, Y. Song, Z. Sun, J. Wang, C. Xu, W. Ouyang, What can simple arithmetic operations do for temporal modeling? in *Proceedings of the IEEE/CVF international conference on computer vision*, (2023), pp.13712–13722
222. H. Yao, W. Wu, Z. Li, Side4video: spatial-temporal side network for memory-efficient image-to-video transfer learning. arXiv preprint arXiv:2311.15769 (2023)

223. C. Ryali, Y.-T. Hu, D. Bolya, C. Wei, H. Fan, P.-Y. Huang, V. Aggarwal, A. Chowdhury, O. Poursaeed, J. Hoffman, et al., Hiera: a hierarchical vision transformer without the bells-and-whistles, in *International Conference on Machine Learning*, (2023), pp. 29441–29454
224. R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, L. Yuan, Y.-G. Jiang, Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), pp. 6312–6322
225. H. Cui, R. Huang, R. Zhang, T. Hayama, Dstsa-gcn: advancing skeleton-based gesture recognition with semantic-aware spatio-temporal topology modeling. Neurocomputing **637**, 130066 (2025)
226. H. Cui, T. Hayama, Joint-partition group attention for skeleton-based action recognition. Signal Process. **224**, 109592 (2024)
227. S. Kim, D. Ahn, B.C. Ko, Cross-modal learning with 3d deformable attention for action recognition, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2023), pp. 10265–1027
228. R. Ding, Y. Wen, J. Liu, N. Dai, F. Meng, M. Liu, Integrating human parsing and pose network for human action recognition, in *CAAI International conference on artificial Intelligence*, (2023), pp. 182–194
229. D. Ahn, S. Kim, H. Hong, B. Chul Ko, Star-transformer: A spatio-temporal cross attention transformer for human action recognition, in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (2023), pp.3319–3328
230. X. Bruce, Y. Liu, K.C. Chan, Multimodal fusion via teacher-student network for indoor action recognition. Proc. AAAI Conf. Artif. Intell. **35**, 3199–3207 (2021)
231. I. Elujide, J. Li, A. Shiran, et al. A Real-time object detection for WiFi CSI-based multiple human activity recognition, in *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)* (2023)
232. M. Dang, G. Liu, Q. Xu et al., Multi-object behavior recognition based on object detection for dense crowds. Expert Syst. Appl. **248**, 123397 (2024)
233. B. Du, J. Zhao, M. Cao, et al. Behavior recognition based on improved faster RCNN, in *2021 14th International congress on image and signal processing, biomedical engineering and informatics (CISP-BMEI)*, (2021) pp.1–6
234. T.-S. Hsiao, H.-Y. Hou, P.T. Lin, C.-Y. Chang, Y.-Y. Chen, C.-L. Yang, Integrating yolo and dg-stgcn systems for enhanced human action recognition, in *2024 International conference on advanced robotics and intelligent systems (ARIS)*, (2024), pp.1–5
235. Y. Duan, Z. Li, B. Shi, Multi-target irregular behavior recognition of chemical laboratory personnel based on improved deepsort method. Processes **12**(12), 2796 (2024)
236. N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, in *2017 IEEE International conference on image processing (ICIP)*, (2017), pp.3645–3649
237. T. Zhou, A. Tao, L. Sun et al., Behavior recognition based on the improved density clustering and context-guided BI-LSTM model. Multimed. Tools Appl. **82**(29), 45471–45488 (2023)
238. H.-H. Chang, Y.-H. Chang, Y.-L. Shih, C.-H. Lin, H.-C. Shih, Basketball player action recognition and tracking using r(2+1)d CNN with spatial-temporal features, in *2024 IEEE 13th global conference on consumer electronics (GCCE)*, (2024), pp.388–389
239. W. Choi, S. Savarese, A unified framework for multi-target tracking and collective activity recognition, in *Computer vision-ECCV*. ed. by W. Choi (Springer, Berlin, 2012)
240. M. Khan, J. Ahmad, A.E. Saddik, et al. Drone-HAT: hybrid attention transformer for complex action recognition in drone surveillance videos, in *2024 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)* (2024)
241. Z. Zhang, H. Lai, D. Huang et al., Reta: 4d radar-based end-to-end joint tracking and activity estimation for low-observable pedestrian safety in cluttered traffic scenarios. IEEE Trans. Intell. Transp. Syst. **25**(5), 4413–4426 (2024)
242. H. Wang, Z.M. Wang, Z.H. Miao, et al. The application of centroid tracking algorithm in video action recognition, in *2021 40th Chinese Control Conference (CCC)* (2021)
243. S. Gupta, J. Malik, Visual semantic role labeling. ArXiv. https://arxiv.org/abs/1505.04474 (2015)
244. M. Nekoui, F.O.T. Cruz, L. Cheng, FALCONS: FAst Learner-grader for CONtorted poses in Sports, in *2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)* (2020)
245. J. Wang, K. Sun, T. Cheng et al., Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **43**(10), 3349–3364 (2021)
246. T.-Y. Pan, W.-L. Tsai, C.-Y. Chang et al., A hierarchical hand gesture recognition framework for sports referee training-based EMG and accelerometer sensors. IEEE Trans. Cybern. **52**, 3172–3183 (2020)
247. T. Nakano, A. Sakata, A. Kishimoto, Estimating blink probability for highlight detection in figure skating videos. ArXiv:2007.01089 (2020)
248. L. Tian, X. Cheng, M. Honda, et al. Multi-technology correction based 3d human pose estimation for jump analysis in figure skating, in *Proceedings* (2020)
249. M. Fani, H. Neher, D.A. Clausi, et al., Hockey action recognition via integrated stacked hourglass network, in *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)* (2017)
250. J. Wang, K. Qiu, H. Peng, et al. AI Coach: deep human pose estimation and analysis for personalized athletic training assistance, in *Proceedings of the 27th ACM international conference on multimedia* (2019)
251. X. Pan, Q. Liu, F. Luan, et al. A study of intelligent rehabilitation robot imitation of human behavior based on kinect, in *2021 IEEE conference on telecommunications, optics and computer science (TOCS)* (2021)
252. W. Wang, J. Li, J. Wang, et al. Multi-sensor patient behavior recognition based on lower limb rehabilitation robot, in *2022 IEEE international conference on mechatronics and automation (ICMA)*, (2022), pp.1444–1451
253. S. Yan, Y. Teng, J.S. Smith, et al. Driver behavior recognition based on deep convolutional neural networks, in *2016 12th International conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD)* (2016)
254. Y. Zhao, T. Li, Y. Dong, A wearable acoustic sensor based driver distraction behaviour recognition, in *2021 International conference on high performance big data and intelligent systems (HPBD &IS)* (2021)
255. J. Seong, C. Lee, D.S. Han, Neural architecture search for real-time driver behavior recognition, in *2022 International conference on artificial intelligence in information and communication (ICAIIC)* (2022)

256. Y. Ban, T. Yamaguchi, H. Okuda, et al. Analysis of SmartphoneWalking behavior of pedestrians at unsignalized intersection based on decisions and motion, in *2023 62nd Annual conference of the society of instrument and control engineers (SICE)* (2023)
257. I. Akhter, M. Javeed, Pedestrian Behavior Recognition via a Smart Graph-based Optimization. 2022 *19th International Bhurban conference on applied sciences and technology (IBCAST)* (2022)
258. S.M. Pang, J.X. Cao, M.Y. Jian et al., Br-gan: a pedestrian trajectory prediction model combined with behavior recognition. IEEE Trans. Intell. Transp. Syst. **23**(12), 24609–24620 (2022)
259. P. Parmar, B.T. Morris, Learning to score olympic events. In: *2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, (2017), pp.76–84
260. C.H. Zhao, B.L. Zhang, J. He, Lian: Recognition of driving postures by contourlet transform and random forests. IET Intell. Transp. Syst. **6**, 161 (2012)
261. R. Mehran, A. Oyama, M. Shah, Abnormal crowd behavior detection using social force model, in 2009 IEEE conference on computer vision and pattern recognition, (2009), pp. 935–942
262. S. Pellegrini, A. Ess, K. Schindler, L. Gool, You'll never walk alone: Modeling social behavior for multi-target tracking. in *2009 IEEE 12th international conference on computer vision*, (2009), pp. 261–268
263. X. Liu, G. Yuan, R. Bing, et al. When skeleton meets motion: adaptive multimodal graph representation fusion for action recognition, in *2024 IEEE international conference on multimedia and expo (ICME)* (2024)
264. W. Wang, D. Tran, M. Feiszli, What makes training multi-modal classification networks hard? in *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (2020), pp.12692–12702
265. H. Fan, S. Liu, Z. Que et al., High-performance acceleration of 2-d and 3-d CNNS on FPGAS using static block floating point. IEEE Transact. Neural Netw. Learn. Syst. **34**, 4473–4487 (2021)
266. K. Mao, P. Jin, Y. Ping et al., Modeling multi-scale sub-group context for group activity recognition. Appl. Intell. **53**, 1149–1161 (2022)
267. X. Zhu, D. Wang, Y. Zhou, Hierarchical spatial-temporal transformer with motion trajectory for individual action and group activity recognition, in: *ICASSP 2023 – 2023 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, (2023), pp.1–5