




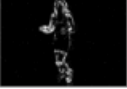

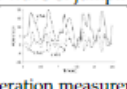
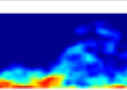
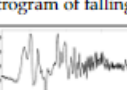


# Human Action Recognition from Various Data Modalities: A Review

**Human Action Recognition (HAR)** is crucial for various real-world applications, including visual surveillance systems, autonomous navigation systems, video retrieval, human-robot interaction, and entertainment. Initially, works focused on RGB or gray-scale videos, but recent years have seen the emergence of other data modalities like skeleton, depth, infrared sequence, point cloud, event stream, audio, acceleration, radar, and WiFi. Non-visual modalities, such as audio, acceleration, radar, and WiFi, are not visually intuitive but can be used for HAR in scenarios requiring privacy protection. These modalities can be used in temporal sequences, fine-grained HAR, and through-wall HAR.

This survey reviews existing deep learning methods for HAR from various data modalities, including RGB, depth, skeleton, infrared sequence, point cloud, event stream, audio, acceleration, radar, and WiFi. It categorizes multi-modality-based HAR methods into fusion-based approaches and cross-modality co-learning-based approaches. The review focuses on recent and advanced deep learning methods for HAR, providing readers with state-of-the-art approaches. The survey also provides comprehensive comparisons of existing methods and their performance on several benchmark datasets, with brief summaries and insightful discussions. This is the first comprehensive review of HAR methods from various data modalities.

Action samples of different data modalities (with pros and cons).

Modality	Example	Pros	Cons
Visual Modality	RGB  Hand-waving	<ul style="list-style-type: none"> <li>Provide rich appearance information</li> <li>Easy to obtain and operate</li> <li>Wide range of applications</li> </ul>	<ul style="list-style-type: none"> <li>Sensitive to viewpoint</li> <li>Sensitive to background</li> <li>Sensitive to illumination</li> </ul>
	3D Skeleton  Looking at watch	<ul style="list-style-type: none"> <li>Provide 3D structural information of subject pose</li> <li>Simple yet informative</li> <li>Insensitive to viewpoint</li> <li>Insensitive to background</li> </ul>	<ul style="list-style-type: none"> <li>Lack of appearance information</li> <li>Lack of detailed shape information</li> <li>Noisy</li> </ul>
	Depth  Mopping floor	<ul style="list-style-type: none"> <li>Provide 3D structural information</li> <li>Provide geometric shape information</li> </ul>	<ul style="list-style-type: none"> <li>Lack of color and texture information</li> <li>Limited workable distance</li> </ul>
	Infrared Sequence  Pushing	<ul style="list-style-type: none"> <li>Workable in dark environments</li> </ul>	<ul style="list-style-type: none"> <li>Lack of color and texture information</li> <li>Susceptible to sunlight</li> </ul>
	Point Cloud  Bending over	<ul style="list-style-type: none"> <li>Provide 3D information</li> <li>Provide geometric shape information</li> <li>Insensitive to viewpoint</li> </ul>	<ul style="list-style-type: none"> <li>Lack of color and texture information</li> <li>High computational complexity</li> </ul>
	Event Stream  Running	<ul style="list-style-type: none"> <li>Avoid much visual redundancy</li> <li>High dynamic range</li> <li>No motion blur</li> </ul>	<ul style="list-style-type: none"> <li>Asynchronous output</li> <li>Spatio-temporally sparse</li> <li>Capturing device is relatively expensive</li> </ul>
Non-visual Modality	Audio  Audio wave of jumping	<ul style="list-style-type: none"> <li>Easy to locate actions in temporal sequence</li> </ul>	<ul style="list-style-type: none"> <li>Lack of appearance information</li> </ul>
	Acceleration  Acceleration measurements of walking	<ul style="list-style-type: none"> <li>Can be used for fine-grained HAR</li> <li>Privacy protecting</li> <li>Low cost</li> </ul>	<ul style="list-style-type: none"> <li>Lack of appearance information</li> <li>Capturing device needs to be carried by subject</li> </ul>
	Radar  Spectrogram of falling	<ul style="list-style-type: none"> <li>Can be used for through-wall HAR</li> <li>Insensitive to illumination</li> <li>Insensitive to weather</li> <li>Privacy protecting</li> </ul>	<ul style="list-style-type: none"> <li>Lack of appearance information</li> <li>Capturing device is relatively expensive</li> </ul>
	WiFi  CSI waveform of falling	<ul style="list-style-type: none"> <li>Simple and convenient</li> <li>Privacy protecting</li> <li>Low cost</li> </ul>	<ul style="list-style-type: none"> <li>Lack of appearance information</li> <li>Sensitive to environments</li> <li>Noisy</li> </ul>

# 1. Single Modality

## 1.1. RGB Modality

RGB modality refers to images or videos captured by RGB cameras, aiming to recreate human eyesight. RGB-based HAR has applications in visual surveillance, autonomous navigation, and sport analysis. However, action recognition from RGB data is challenging due to variations in backgrounds, viewpoints, scales, and illumination conditions. Most existing works focus on using videos for HAR, with only a few using static images. Advanced deep learning works for RGB-based are divided into four categories: two-stream 2D Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), 3D CNN, and Transformer-based methods. These methods have been developed due to their strong representation capability and superior performance.

### 1.1.1. Two-Stream 2D CNN-Based Method

The two-stream **2D CNN** (Convolution **N**eural **N**etwork) framework consists of two 2D CNN branches extracting different input features from RGB videos for HAR. The final result is obtained through fusion strategies. Classic two-stream methods include Simonyan and Zisserman's model, which consists of a spatial and temporal network. Karpathy et al. used low-resolution RGB frames and high-resolution center crops to speed up computation. Wang et al. extracted convolutional feature maps from multi-scale video frames and optical flows, and C  ron et al. used human body joints to crop multiple parts from RGB and optical flow images.

Some researchers have proposed methods to reduce computational costs associated with accurate optical flow training. Zhang et al. proposed a teacher-student framework, transferring knowledge from a teacher network trained on optical flow data to a student network trained on motion vectors from compressed videos. Piergiovanni and Ryoo proposed a trainable flow layer, capturing motion information without computing optical flows. Other researchers have extended two-stream CNN architectures, recursively predicted frame discriminative importance, and proposed super-resolution methods for low-resolution videos. Fusion strategies have been studied to reduce parameters while maintaining accuracy.

### 1.1.2. RNN-Based Methods

RNNs are useful for analyzing temporal data due to their recurrent connections. However, traditional RNNs struggle with long-term temporal dependency, leading to the adoption of gated RNN architectures like Long-Short Term Memory (LSTM) for modeling long-term temporal dynamics in video sequences. RNN-based methods typically use 2D CNNs as feature extractors and LSTM models for HAR. Examples include the Long-term Recurrent Convolutional Network (LRCN), stacked LSTM frameworks, encoder LSTMs, C2LSTMs, and Bi-directional LSTMs for efficient HAR.

### 1.1.3. 3D CNN-Based Methods

Research has extended 2D CNNs to 3D structures to model spatial and temporal context information in videos, crucial for HAR. Early works segmented human subjects in videos and fed them to a 3D CNN model. Tran et al. introduced a 3D CNN model called C3D, but these networks were mainly used for clip-level learning. Several approaches focused on modeling long-range spatio-temporal dependencies in videos. Diba et al. extended DenseNet with 3D filters and pooling kernels, designed a Temporal 3D CNN (T3D), and introduced a new block embedded in architectures like ResNext and ResNet. Long-term Temporal Convolution (LTC) frames, Timeception, non-local operations, and Channel Independent Directional Convolution (CIDC) have been proposed to better capture long-term temporal dynamics.

Several studies have investigated 3D CNN models to improve the performance of HAR (Human Representation Analysis) in various applications. These include the two-stream Inflated 3D CNN (I3D), which inflates the convolutional and pooling kernels of a 2D CNN with an additional temporal dimension. Other works have integrated a two-stream 3D CNN with an LSTM model to capture long-range temporal dependencies. The ECO architecture uses 2D CNNs to extract spatial features, which are then fed to 3D CNNs to model long-term dependencies for HAR.

Additionally, deep learning frameworks combining 2D CNNs and 3D CNNs have been investigated for HAR. For example, the ECO architecture uses 2D CNNs to extract spatial features, which are then fed to 3D CNNs to model long-term dependencies for HAR.

Some works have also addressed other problems in 3D CNNs, such as spatio-temporal fusion, viewpoint variation, and knowledge distillation. Stroud et al. introduced a Distilled 3D Network (D3D) consisting of a student and teacher network, which distilled knowledge from a teacher network trained on optical flow sequences. Shou et al. proposed an adversarial framework with a lightweight generator to approximate flow information by refining the noisy and coarse motion vector available in the compressed video. Wang et al. adopted an efficient learnable correlation operator to better learn motion information from 3D appearance features.

## 1.2. Skeleton Modality

Skeleton sequences, which encode human body joints, are suitable for Human-Robot (HAR) due to their informative human motions. Skeleton data can be obtained through pose estimation algorithms on RGB videos, depth maps, or motion capture systems. However, motion capture systems are insensitive to view and lighting, making skeleton data less convenient. Recent works on HAR have used skeleton data from depth maps or RGB videos. Skeleton data offers advantages such as body structure and pose information, simple representation, scale invariance, and robustness against clothing textures and backgrounds.

Due to its advantages and availability of accurate and low-cost depth sensors, skeleton-based HAR has gained attention in the research community. Early works focused on extracting hand-crafted spatial and temporal features from skeleton sequences. Deep learning has become the mainstream research in this field due to its strong feature learning capability.

### **1.2.1. RNN-Based Methods**

RNNs and LSTMs have been used to learn dynamic dependencies in sequential data, enabling various methods to model temporal context information within skeleton sequences for HAR. Classical methods include end-to-end hierarchical RNNs, which divide the human skeleton into five body parts and feed them to multiple bidirectional RNNs. Differential RNNs learn salient spatio-temporal information by quantifying the change of information gain caused by salient motions between frames. The Derivative of States (DoS) is proposed inside the LSTM unit to control the information flow over time. Zhu et al. introduced a novel mechanism for automatic co-occurrence mining, while Sharoudy et al. proposed a Part-aware LSTM (P-LSTM) to simulate relations among different body parts. Liu et al. extended RNN design to both temporal and spatial domains, using tree structure-based skeleton traversal methods and trust gates to deal with noise and occlusions. Global Context-Aware Attention LSTM (GCA-LSTM) selectively focuses on informative joints using global context information. Two-stream RNN structures, deep LSTM with spatio-temporal attention, and ensemble Temporal Sliding LSTM frameworks have been proposed to model variable temporal dynamics of skeleton sequences.

### **1.2.2. CNN-Based Methods**

CNNs have been successful in 2D image analysis due to their ability to learn features in the spatial domain. However, modeling spatio-temporal information in skeleton-based HAR poses a challenge. Advanced approaches have been proposed, such as applying temporal convolution on skeleton data or representing skeleton sequences as pseudo-images that are fed to standard CNNs for HAR. These pseudo-images encode spatial structure information in each frame and temporal dynamic information between frames.

Hou et al. and Wang et al. proposed skeleton optical spectra and joint trajectory maps, which encoded spatio-temporal information into color-texture images and adopted CNNs for HAR. The Joint Distance Map (JDM) produces view-invariant color-texture images encoding pair-wise distances of skeleton joints. Ke et al. transformed each skeleton sequence into three "video clips" and fed them to a pre-trained CNN for HAR. Kim and Reiter used the Temporal CNN (TCN) for interpretable spatio-temporal representations. Caetano et al. introduced SkeleMotion and the Tree Structure Reference Joints Image (TSRJI) as representations of skeleton sequences for HAR.

Several researchers have focused on addressing specific problems, such as viewpoint variation and features learned from skeleton data that are not always translation, scale, or rotation invariant.

### **1.3. Depth Modality**

Depth maps are images that represent the distance information from a given viewpoint to points in the scene. They provide reliable 3D structural and geometric shape information of human subjects, making them useful for Human-Robot (HAR) recognition. Different types of devices have been developed to obtain depth images, including active sensors like Time-of-Flight and structured-light-based cameras and passive sensors like stereo cameras. Active sensors emit radiation and measure the reflected energy from objects to acquire depth information, while passive sensors measure natural energy emitted or reflected by objects in the scene.

Passive depth map generation is usually computationally expensive and can be ineffective in texture-less regions or highly textured regions with repetitive patterns. In this section, we review methods that used depth maps for HAR, focusing on methods using depth videos captured by active sensors. Deep learning models have shown to be more powerful and achieved better performance for HAR from depth maps. Examples include a deep learning framework using weighted hierarchical DMMs, representing depth sequences with three pairs of structured dynamic images at the body, body part, and joint levels, feeding them to CNNs followed by a score fusion module for fine-grained HAR, and transferring human data obtained from different views to a view-invariant high-level space.

In addition to depth maps obtained with active sensors or stereo cameras, there has been another approach designed for depth-based HAR from RGB videos. In general, the depth modality provides geometric shape information that is useful for HAR, but depth data is often not used alone due to the lack of appearance information.

### **1.4. Infrared Modality**

Infrared sensors are ideal for night vision (HAR) due to their lack of reliance on external ambient light. These sensors can be active or passive, with some like Kinect relying on active infrared technology to detect objects in the scene. Thermal sensors, on the other hand, work by detecting heat energy emitted from targets.

Several deep learning methods have been proposed for HAR from infrared data. For instance, cropping low-resolution thermal images by the gravity center of human regions and passing them through a CNN followed by an LSTM layer to model spatio-temporal information

for HAR. Shah et al. achieved real-time HAR using a 3D CNN, while Megloulou et al. passed optical flow information computed from thermal sequences into a 3D

Multi-stream architectures have also been proposed, including two-stream CNN models, three-stream CNNs, and four-stream architectures. These models capture complementary information on appearance from still thermal frames and motion between frames. Imran and Raman proposed a four-stream architecture, where each stream consists of a CNN followed by an LSTM.

### **1.5. Point Cloud Modality**

Point cloud data is a 3D data modality that represents the spatial distribution and surface characteristics of a target under a spatial reference system. It can be obtained through 3D sensors like LiDAR and Kinect or image-based 3D reconstruction. Point cloud data can be used for HAR, as it can represent spatial silhouettes and geometric shapes of subjects. Early methods focused on extracting hand-crafted spatio-temporal descriptors from point cloud sequences, but current research focuses on deep learning architectures that generally achieve better performance.

MeteorNet, a self-supervised learning framework, directly stacks multi-frame point clouds and calculates local features by aggregating information from spatio-temporal neighboring points. PSTNet disentangles space and time to reduce the impacts of spatial irregularity of points on temporal modeling. Wang et al. proposed an anchor-based spatio-temporal attention convolution model to capture the dynamics of 3D point cloud sequences, but these methods are not able to sufficiently capture long-term relationships within point cloud sequences.

Transformers have gained increasing attention in point cloud-based HAR, such as Point Attention Transformer (PAT), Point 4D convolution, and Point Spatial-Temporal Transformer (PST2). A self-attention-based module, Spatio-Temporal Self-Attention (STSA), was introduced to capture spatial-temporal context information for action recognition in 3D point clouds.

### **1.6. Event Stream Modality**

Event cameras, also known as neuromorphic cameras or dynamic-vision sensors, have gained attention for their ability to capture illumination changes and produce asynchronous events independently for each pixel. These cameras are suitable for HAR due to their high dynamic range, low latency, low power consumption, and no motion blur. However, the information obtained with event cameras is generally spatio-temporally sparse and asynchronous. Common event cameras include the Dynamic Vision Sensor (DVS) and the Dynamic and Active-pixel Vision Sensor (DAVIS).

Some existing methods focus on designing event aggregation strategies, which convert the asynchronous output of the event camera into synchronous visual frames that can be

processed with conventional computer vision techniques. Deep learning methods have been more popular recently, with methods such as Innocenti et al. building a sequence of binary representations from the raw event data, Huang et al. using time-stamp image encoding, and Bi et al. representing events as graphs and using a GCN network for end-to-end feature learning directly from the raw event data.

In general, the event stream modality is an emerging modality for HAR, as processing event data is computationally cheap and does not usually contain background information, which can be helpful for action understanding. However, event stream data cannot generally be effectively and directly processable using conventional video analysis techniques, making it a challenging research problem.

### **1.7. Audio Modality**

Audio signals, often found in videos, can be used for HAR to locate actions, reducing human labeling efforts and computational costs. Some deep learning-based methods have been proposed for general activity recognition from audio signals, but only a few have been proposed for HAR from audio signals alone. Audio modality alone is not popular due to its insufficient information for accurate HAR, but it can serve as complementary information for more reliable and efficient HAR. Most audio-based HAR methods focus on multi-modal deep learning techniques.

### **1.8. Acceleration Modality**

Acceleration signals from accelerometers are used for HAR due to their robustness against occlusion, viewpoint, lighting, and background variations. Tri-axial accelerometers can provide estimates of acceleration along x, y, and z axes for human activity analysis. Although the human body size and proportion vary, the acceleration signal usually does not have obvious intra-class variations for the same action. This method achieves high accuracy and is adopted for remote monitoring systems while addressing privacy issues. Deep learning networks have been widely used for acceleration-based HAR, with various techniques being proposed to extract spatial and temporal features from raw acceleration data. Some acceleration-based methods focus on fall detection tasks. Despite its privacy-protecting characteristics, acceleration is often used for fine-grained HAR and elderly care due to its privacy-protecting nature. However, wearable sensors and sensor position can affect HAR performance.

### **1.9. Radar Modality**

Radar technology, such as Doppler and FMCW radars, is commonly used for HAR (Human-Robotic Systems) due to their ability to detect radial velocity and frequency changes based on distance. These micro-Doppler signatures contain the target's motion and structure information, making them suitable for HAR. Spectrograms from radars offer advantages such

as robustness to illumination and weather conditions, privacy protection, and through-wall HAR capabilities.

Recently, several deep learning architectures have been proposed to predict action classes in various scenarios, including aquatic activities, geometrical locations, and simulated environments. Two-stream architectures have also been investigated, with some interpreting micro-Doppler spectrograms as temporal sequences. RNN-based architectures have also been proposed, using LSTM models and stacked RNN models to predict action classes.

Despite the advantages of radar for HAR, radars are relatively expensive. While some datasets have shown satisfactory results, there is still room for further development in radar-based methods. Future directions in this area include handling more complex actions in real-world scenarios using radar data.

### **1.10. Wifi Modality**

Radar technology, such as Doppler and FMCW radars, is commonly used for HAR (Human-Robotic Systems) due to their ability to detect radial velocity and frequency changes. These spectrograms can provide motion and structure information for HAR, making them suitable for various scenarios. Recent research has proposed several deep learning architectures, including micro-Doppler spectrogram images, two-stream architectures, and RNN-based architectures.

The radar modality has advantages, such as being suitable for some scenarios, but it is relatively expensive. While radar data has achieved satisfactory results, there is still room for development in radar-based methods. Future directions in this area include handling more complex actions in real-world scenarios with radar data.

The spatial features of the CSI signal can be extracted from a fully connected layer of a pre-trained CNN, fed to a Bi-LSTM to capture temporal information for HAR. Gao et al. transformed the CSI signal into radio images, which were then fed to a deep sparse auto-encoder to learn discriminative features for HAR.

The WiFi modality can be used for HAR in some scenarios due to its convenience. However, challenges such as effectively using CSI phase and amplitude information and improving robustness in dynamic environments need to be addressed. Other modalities, such as radio frequency, energy harvester, gyroscope, electromyography, and pressure, have also been used for HAR.

## **2. Multi-Modality**

Multi-modal machine learning is a modeling approach that processes and correlates sensory information from multiple modalities, providing robust and accurate HAR. It combines the advantages of various data modalities, with two main types: fusion and



co-learning. Fusion integrates information from multiple modalities for training and inference, while co-learning transfers knowledge between different data modalities.

## **2.1. Fusion**

Multi-modality fusion is a technique used in HAR to enhance performance by combining the strengths of different data modalities. Two commonly used multi-modality fusion schemes are score fusion and feature fusion. Score fusion integrates decisions made by different modalities, while feature fusion combines features from different modalities to yield aggregated features that are highly discriminative and powerful for HAR. Data fusion, combining multi-modality input data before feature extraction, has also been utilized.

### **2.1.1. Fusion of Visual Modalities**

The emergence of low-cost RGB-D cameras has led to the development of multi-modality datasets, leading to the development of multi-modality fusion-based HAR methods. These methods focus on the fusion of visual modalities, such as RGB and depth data. RGB and depth videos capture rich appearance and 3D shape information, which are complementary and can be used for HAR. Early methods focused on extracting hand-crafted features that capture spatio-temporal structural relationships from the RGB and depth modalities.

Several deep learning methods have been introduced to fuse RGB and depth data modalities. For example, a four-stream deep CNN was introduced to extract features from different representations of depth data and RGB data. Wang et al. extracted scene flow features from spatially aligned and temporally synchronized RGB and depth frames, and their classification scores were fused to perform HAR.

Transformers have also become popular for multi-modal fusion. Zhou et al. proposed a decoupling and recoupling spatio-temporal representation approach, while Dhiman et al. designed a two-stream network composed of a motion stream and a Shape Temporal Dynamic (STD) stream to encode features from RGB and depth videos.

Fusion of RGB and skeleton data is another area of research. Zhao et al. introduced a two-stream deep network, which consists of an RNN and a CNN to process skeleton and RGB data. Liu et al. proposed a spatio-temporal LSTM network, and Song et al. proposed a learning framework containing two streams of skeleton-guided deep CNN to extract features from RGB and optical flow.

Finally, Cai et al. introduced a two-stream GCN network, RGBPose-Conv3D, and a Two-Pathway Vision Transformer (TP-ViT) to fuse RGB and skeleton data for more accurate HAR.

### **2.1.2. Fusion of Visual and Non-visual Modalities**

Visual and non-visual modalities can be fused to enhance the accuracy and robustness of Human-Robot (HAR) models. Audio data provides complementary information to visual data, and several deep learning-based methods have been proposed to fuse these two types of modalities for HAR. Wang et al. introduced a three-stream CNN to extract multi-modal features from audio signal, RGB frames, and optical flows, while Owens and Efros trained a two-stream CNN in self-supervised manner to identify any misalignment between the audio and visual sequences.

Kazakos et al. introduced a Temporal Binding Network (TBN), which takes audio, RGB, and optical flow as inputs for egocentric HAR. Gao et al. used audio signal to reduce temporal redundancies in videos, distilling knowledge from a teacher network trained on video clips to a student network trained on image-audio pairs for efficient HAR.

Alfasly et al. adopted a transformer model, BERT, to obtain sentence-based semantic embeddings of each textual label in audio and video datasets. Zhang et al. proposed an audio-adaptive model leveraging the rich audio information to adjust the visual representation, along with an audio-infused recognizer to maintain domain-irrelevant features.

Dawar et al. represented inertial signal as an image, and two CNNs were used to fuse depth images and inertial signals using score fusion. Wei et al. fed 3D video frames and 2D inertial images to a 3D CNN and a 2D CNN for HAR, and the scores of the two models were fused during testing for better classification.

However, the task of effective modality fusion remains largely open, as most existing multi-modality methods have complicated architectures that require high computational costs.

## **2.2. CO-Learning**

Co-learning involves transferring knowledge from auxiliary modalities to improve model learning on another modality. This method overcomes single data limitations and enhances performance by utilizing data from auxiliary modalities during training and testing, especially for fewer samples.

### **2.2.1. Co-Learning with Visual Modalities**

Co-learning-based HAR methods primarily focus on visual modalities, such as RGB with depth and RGB with skeleton modalities. Knowledge transfer between these modalities improves the representation capability of each modality, especially when one has limited annotated data for training. Some methods use hand-crafted features for co-learning. Garcia et al. proposed several deep learning-based HAR methods based on cross-modality knowledge distillation. They used a knowledge distillation framework to distill knowledge from a teacher network taking depth videos to a RGB-based student network. Another study used a CNN+LSTM network to perform classification based on RGB videos, and an LSTM

model trained on skeleton data to act as a regulator. Hong et al. used the Video Pose Distillation strategy in a teacher-student architecture. Other visual modalities have also been investigated, such as a transferable generative model that generates fake feature representations of RGB videos and the PIX2NVS emulator embedded into a teacher-student framework.

### **2.2.2. Co-Learning with Visual and Non-visual Modalities**

Numerous studies have explored co-learning between visual and non-visual modalities, such as training teacher networks on non-visual modalities like acceleration, gyroscope, and orientation signal. These networks were trained using knowledge distillation with an attention mechanism, integrating classification scores and attention weights. Another study proposed a knowledge distillation framework to transfer knowledge from a teacher network trained on RGB videos to a student network using raw sound data. Yang et al. proposed a Cross-modal Interactive Alignment model for unsupervised domain adaptive video action recognition. Some works also leveraged the correlation between different modalities for self-supervised learning, such as unsupervised clustering in audio/video modality and audio-visual correspondence and temporal synchronization. Other works used narrations of videos as a weak supervisory signal to jointly learn video and text representations for action recognition and detection.

## **3. Conclusion**

This paper provides a comprehensive review of HAR methods using various data modalities and surveys multi-modality recognition methods, including Fusion and Co-learning methods, which have gained significant research attention in recent decades.