



## Article

# RGB-D Data-based Action Recognition: A Review

Muhammad Bilal Shaikh <sup>1,†,‡</sup>  and Douglas Chai <sup>1,‡</sup> <sup>1</sup> School of Engineering, Edith Cowan University, Perth, WA 6027, Australia; m.shaikh@ecu.edu.au

\* Correspondence: m.shaikh@ecu.edu.au;

‡ These authors contributed equally to this work.

Version January 18, 2021 submitted to Journal Not Specified

**Abstract:** Classification of human actions from uni-modal and multi-modal datasets is an ongoing research problem in computer vision. This review is aimed to scope current literature on data-fusion and action-recognition techniques and to identify gaps and future research direction. Success in producing cost-effective and portable vision-based sensors has dramatically increased the number and size of datasets. The rise in number of action recognition datasets intersects with advances in deep-learning architectures and computational support, both of which offer significant research opportunities. Naturally, each action-data modality - such as RGB, depth, skeleton, and infrared - has distinct characteristics; therefore, it is important to exploit the value of each modality for better action recognition. In this article we will focus solely on areas such as data fusion and recognition techniques in the context of vision with a uni-modal and multi-modal perspective. We conclude by discussing research challenges, emerging trends, and possible future research directions.

**Keywords:** Action Recognition; Deep Learning; Data Fusion

## 1. Introduction

Human action recognition has recently gained increasing attention from computer vision researchers with applications in robot vision, multimedia content search, video surveillance, and motion tracking systems. The recent developments in artificial intelligence have stimulated computer vision researchers to investigate problems in recognizing actions. Coupled with the vast amount of digital data available today, the rise of deep learning has resulted in a dramatic increase in computing resources and offers attractive opportunities for designing efficient action recognition systems.

The development of low-cost sensors like ASUS Xtion [1], Microsoft Kinect [2] and Intel RealSense [3] has sparked further research into action recognition. These sensors collect data in various modalities such as RGB video, depth, skeleton, and infrared (IR). All of these modalities have their own characteristics that can help answer challenges related to action data and provide potential opportunities for computer vision researchers to examine vision data from different perspectives.

Herath *et al.* [4] have defined action as “the most elementary human-surrounding interaction with a meaning.” Human Action Recognition is the labeling or annotation of actions performed by humans within a given video, where it becomes the classification of goals of a human agent in a series of image frames. From a temporal perspective, action recognition is the localization of action in a video, where a sequence of image frames is related to a particular class of action. Due to their multi-faceted nature, some of these approaches refer to action recognition as plan recognition, goal recognition, intent recognition, behavior recognition, location estimation, event recognition, action recognition, and interaction recognition. Some of the terms referenced in the literature in relation to action are defined in Table 1.

Early research on Human Action Recognition was dominated by the analysis of still images or videos [5–10], localizing the actor in a video spatio-temporally using bounding boxes, temporal

**Table 1.** Terms related to uni-modal and multi-modal action recognition

Term	Definition
Gesture, Mime, Sign	Basic movement or positioning of the hand, arm, body, or head that communicates an idea, emotion, etc.
Action, Event	A type of motion performed by a single person during short time period and involves multiple body parts.
Activity	Composed of a sequence of actions.
Interaction	A type of motion performed by two actors; one actor is human while the other may be human or an object.
Uni-modal, Single-mode	having or involving one mode.
Multi-modal, Multi-type, Multi-format	Different types of data acquired through sensors.
Fusion, Mixture, Combination	A process for combining different types of sensor data.

extent, and a spatio-temporal cuboid which contains a particular action. Action recognition remains challenging due to problems posed by background clutter, partial occlusion, viewpoint, lighting changes, execution rate, and biometric variation. These challenges remain even with the application of current deep-learning-based approaches [4,11]. Understanding information from images is a challenging process that has engaged thousands of researchers for over four decades and studies are still far from developing a general-purpose machine that can “see” [12].

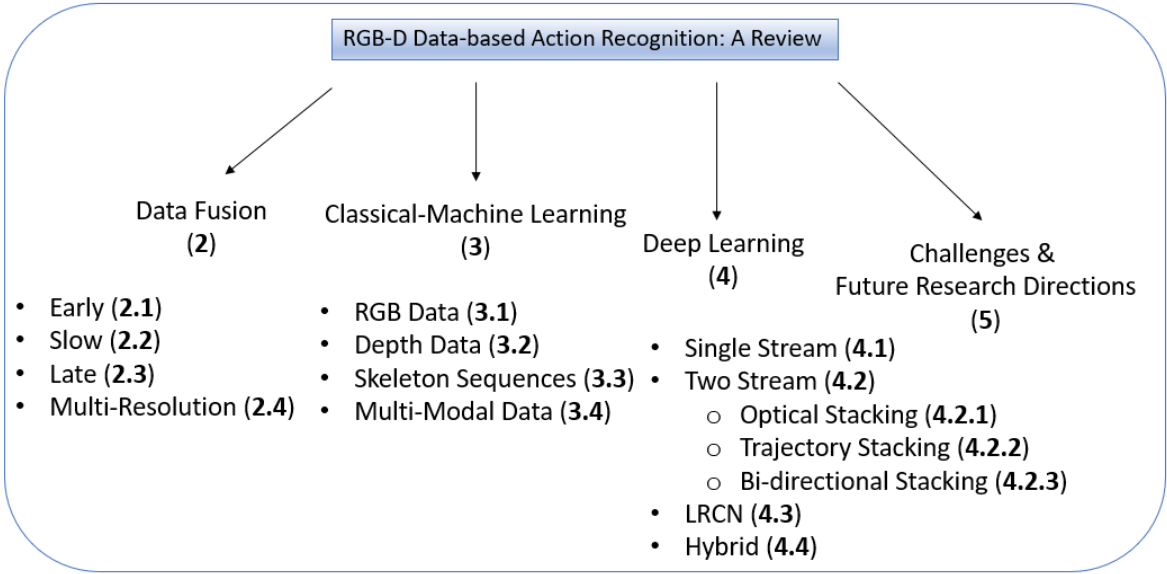
Human Action Recognition has many applications, including the automated annotation of user videos; indexing and retrieving user videos; automated surveillance; monitoring elderly patients using specially adapted cameras; robot operations; and live blogging of actions. In recent times, the availability of massive amounts of video data has provided significance to the understanding of video data (through a sequence of images) with the possibility of solving problems such as scene identification, searching through video content, and interaction recognition through video scenes [13].

Several survey papers [14–24] have discussed action recognition from different perspectives. To our knowledge, there is no review with a focus specifically on data-fusion and vision-based action recognition in the context of uni-modal and multi-modal data. The literature for paper has been selected by using the keywords ‘action recognition’ and ‘multi-modal action recognition’ over the period from 2010 to 2020. In vision-based action recognition, classification techniques takes distinctive characteristics from each modality and apply computer vision methods. This article offer computer vision researchers a set of potential opportunities to explore vision data by exploiting natural characteristics of different modalities.

A novel contribution of this review is the focus on different uni-modal and multi-modal architectures of deep learning-based methods used in action recognition. Apart from this, this work distinguishes itself from other studies through the following contributions:

1. Review of state-of-the-art action recognition techniques on common uni-modal and multi-modal datasets that will provide readers with an overview of recent developments in action recognition.
2. Analysis of current methods from a perspective of multi-modality and hybrid classification methods.
3. Intuitive categorization and analysis of recent and advanced classical machine learning-based and deep learning-based techniques.
4. Discussion of the challenges of data fusion and action recognition and potentials future research directions.

The rest of the article is organized as follows: Section 2 discusses the use of different data-fusion techniques used in action recognition systems. Section 3 reviews the classical machine learning-based methods. Section 4 provides a discussion on deep learning-based methods of action recognition.



**Figure 1.** Structure of this paper. Numbers in brackets refers to section heading.

Section 5 outlines different challenges in data fusion, and action recognition techniques and discusses the future research directions. Section 6 concludes this paper. Fig. 1 shows the hierarchical structure of this paper.

**2. Data-Fusion Techniques**

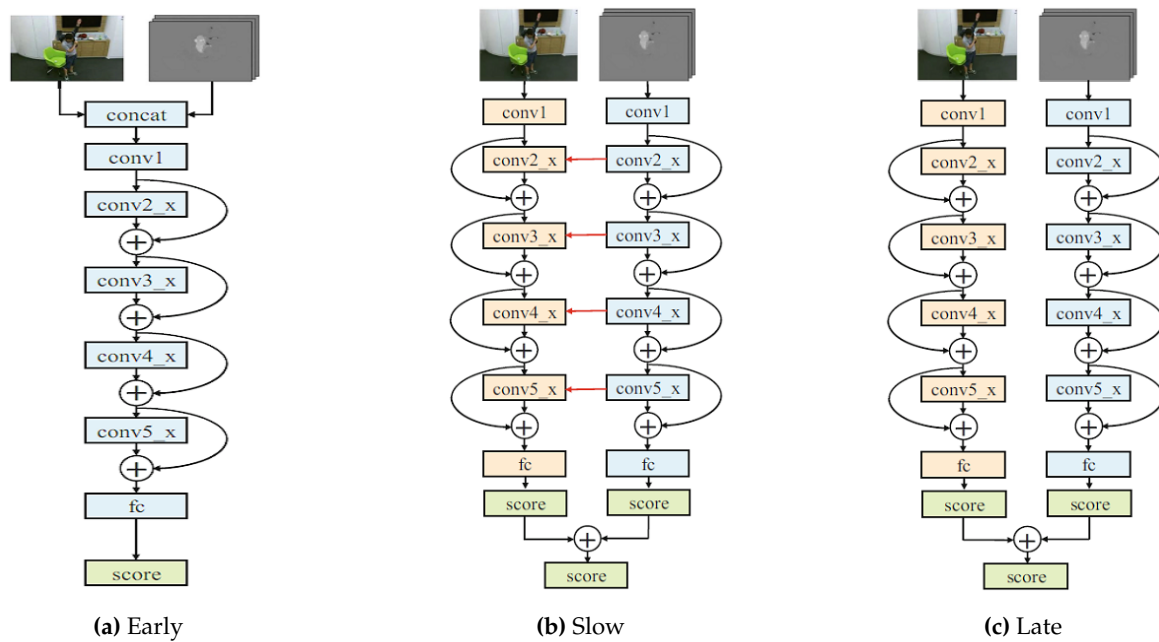
Data fusion supports diversity which enhances the uses, advantages, and analysis of ways that cannot be achieved through a single modality. Fusion techniques can be deployed at different stages in the action recognition process to acquire combinations of distinct information. The following are some popular ways of fusing action data in multi-modal datasets.

*2.1. Early Fusion*

The early fusion approach captures information and combines it across a raw or data level. To achieve this, the filters on the first layer in the neural network are modified. This direct and early connectivity of the raw data helps the network to detect the fused feature vectors at an early stage. For an entire action data sequence, some randomly selected sample instances are often passed through the system, and their class predictions are then averaged to produce action class predictions, as illustrated in Fig. 2.

*2.2. Slow Fusion*

The slow fusion approach fuses the features extracted from raw-data throughout the neural network so that the higher layers have access to more global information. This is achieved by performing a CNN, RNN, or LSTM-based operation to calculate the weights and extend the connectivity of all layers. For example, as shown in Fig. 2, raw-data from two-different modalities is filtered in the first convolution layer. The next layers iterate this process in the network with different filter configurations. Therefore, the information across all the input data can be assessed by the subsequent convolution layers. Given an entire human action data sequence, action classification is performed over the entire multi-modal dataset through the network and then averaging individual predictions throughout the action sequence.



**Figure 2.** An example of Early, Slow and Late fusion in multimodal HAR (taken from [25]). Note that the input modalities are not limited in the above two modalities.

### 2.3. Late Fusion

The late fusion approach combines the action information at the deepest layers in the network. For example, An HAR architecture network consisting of two separate CNN-based networks with shared parameters up to the last convolution layer. The outputs of the last convolution layer of these two separate network streams are processed to the fully connected layer. Global action characteristics are fused and classification score are then averaged or concatenated by different holistic operations at the score layer. This approach have been relatively successful in most of the HAR systems. An illustration of late fusion is shown in Fig. 2.

### 2.4. Multi-resolution

In order to speed up the above-mentioned models while retaining their accuracy, a multi-resolution architecture has been proposed by Karpathy *et al.* [26]. The multi-resolution model consists of two separate networks (fovea and cortex networks) over two spatial resolutions. The architectures of fovea and cortex networks are similar to the single-frame architecture. However, instead of accepting the original input, these networks record reduced sized inputs. More precisely, the input to the fovea model is the center region at the original spatial resolution. In contrast, for the context stream, the down-sampled frames at half the original resolution are used. The total dimensionality of the inputs is, therefore, halved. Moreover, the last pooling layer is removed from both the fovea and cortex networks. The activation outputs of both networks are concatenated and fed into the first fully connected layer.

All the above-mentioned network variations were trained on the Sports-1M dataset [26], which consists of 200,000 test videos. The results showed that the variation among different Convolutional Neural Network (CNN) architectures (e.g., Single Frame, Multi-Resolution, Early, Late, and Slow Fusion) was surprisingly insignificant. Furthermore, the results were significantly worse than the state-of-the-art hand-crafted shallow models. One reason for this may be because these models cannot capture the motion information in many cases. For example, the slow-fusion model is expected to implicitly learn the spatio-temporal features in its first layers, which is a difficult task. To resolve this issue, a two-stream CNNs model was proposed by Simonyan and Zisserman [27] to explicitly take into account both spatial and temporal information in a single end-to-end learning framework.

**Table 2.** Summary of popular action-recognition datasets and methods that achieved the best recognition accuracy. Note that PDF stands for probability distribution function, i3D stands for inflated 3D, and GCN stands for Graph Convolutional Networks.

Year	Ref.	Methods	Action Datasets	Action Datasets																			
				VIRAT Ground 2.0 [28]	ActivityNet [29]	ActionNet-VE [30]	UCF101 [31]	THUMOS'14 [32]	Something-Something v2 [33]	VIVA Hand Gestures [34]	UTD-MHAD [35]	EgoGesture [36]	AVA 2.1 [37]	Something-Something v1 [38]	Charades [39]	NTU RGB+D 120 [40]	miniSports [41]	Sports-1M [26]	IRD [42,43]	HMDB-51 [44]	ICVL-4 [42,43]	NTU RGB+D 60 [45]	Jester [46]
2014	[47]	PDF		66																			
2016	[48]	LSTM			75																		
2017	[49]	Ontology/Rule				90																	
2017	[50]	i3D CNN					98																
2018	[51]	CNN						37															
2018	[33]	2D CNN							66														
2018	[52]	3D CNN								86		93											
2018	[43]	i3D CNN									92										94		
2019	[53]	iD CNN+Attention											27										
2019	[38]	2D CNN												55									
2019	[54]	CNN													56								
2019	[41]	i3D CNN														74							
2019	[55]	3D CNN															75						
2019	[42]	GCN																80		91			
2019	[56]	CNN																	82				
2019	[57]	CNN																					
2020	[58]	3D CNN														85					97		

3. Classical Machine-Learning-Based Techniques

Classical machine-learning-based action recognition techniques use hand-crafted features and can be classified on the basis of RGB data [18], depth data [59,60], skeleton sequences [61], and methods using a combination [62] of these data modalities. Table. 2 summarizes the best performing techniques which achieved benchmark accuracies for popular uni-modal and multi-modal datasets in action recognition research. Classical machine learning-based methods have divided on the basis of their modality into four categories: RGB, depth, skeleton and multi-modal.

3.1. RGB-Data-Based Techniques

Features are extracted to acquire (1) human movements, (2) spatial and temporal changes using Spatio-temporal volume-based methods [63,64], and Spatio-Temporal Interesting Points (STIP)-based methods [65,66], (3) the trajectory of skeleton joints [67,68] using action representation methods, and (4) human image sequences. The hierarchy of handcrafted feature-based techniques that use classical machine learning methods for action recognition is illustrated in Fig. 3.

The spatio-temporal volume-based technique was the earliest one to use a 3D spatio-temporal template matching technique for action recognition. Bobick and Davis [69] have proposed Motion Energy Image (MEI) and Motion History Image (MHI) for the representation of action. Zhang *et al.* [70] have used polar co-ordinates to divide the middle regions of the human body into MHI and Motion-Context descriptors to represent action. Klaser *et al.* [71] have extended the histogram of gradients (HOG) feature of an image to a 3D HOG feature to describe human action, and Somasundaram *et al.* [72] used dictionary-learning methods and sparse representation to measure intra-sequence similarity in spatio-temporal dimensions.

STIP-based methods extract key regions (the position that changes most) from an image sequence to represent action. STIP-based methods are mostly extensions of object recognition methods applied to images; therefore, a key region detection method, feature vector, and a classification algorithm need to be determined. 3D-Harris spatio-temporal points [73,74] are popular. The idea here is to extend local feature detection to the 3D spatio-temporal domain. This leads to the computation of

feature descriptors and visual dictionary is learnt to represent actions. The spatio-temporal attention mechanisms for action representation were proposed by Nguyen *et al.* [63]. Peng *et al.* have [75] proposed a hybrid supervector method for action representation that achieved significant performance on common datasets. Nazir *et al.* [76] have combined the 3D-Harris spatio-temporal features and 3D scale-invariance feature transform detection methods with visual histograms to represent actions.

Trajectory-based features use the tracking path of joints in a human skeleton or key points for representation of actions. Wang *et al.* [68] have proposed improved Dense Trajectories (iDT) which use an optical flow approach to track feature points by sampling dense point clouds. Gaidon *et al.* [77] have proposed the idea of using split clustering to analyze local motion trajectories to represent motion levels of action features. Wang *et al.* [78] proposed the fusion of iDT features and human detection that reduced noise from background trajectories. Stacked Fisher vectors were used by Peng *et al.* [79] that further improved iDT. However, Wang *et al.* [47] have used a variational Bayes method, while Moon *et al.* [49], through an ontology and rule-based methodology, have produced benchmark action recognition techniques on RGB-based datasets.

### 3.2. Depth-Data-Based Techniques

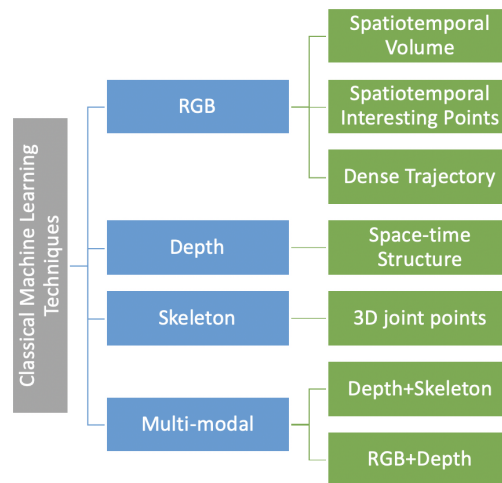
Motion changes in the depth maps of the human body are used to represent action. Depth data can be observed as a space—time structure which is extracted from the appearance and motion information to describe human actions. Yang *et al.* [80] have proposed a supernormal vector feature through a depth-map sequence for action representation. Oreifej *et al.* [59] have proposed an orientation histogram feature of 4D normal vectors to represent the appearance information of a 3D spatio-temporal depth structure. Rehmani *et al.* [81] have proposed the idea of the main direction of a depth-curved surface where a perspective-independent feature and a principal component histogram is used to represent action. Yang *et al.* [82] have proposed the Depth Motion Map (DMM) to project spatio-temporal depth structure onto motion history maps. More recent motion history maps are represented by HOG features in series to represent actions. Chen *et al.* [83] have used local binary features instead of HOG features. Chen *et al.* [84] also investigated spatio-temporal depth structure from front, side, and upper directions. Miao *et al.* [85] have considered discrete cosine variation to compress the depth map and represent action through features using transform coefficients.

### 3.3. Skeleton Sequences-Based Techniques

Changes in position and appearance changes in human joint points between frames are used to describe action. Xia *et al.* [86] have modeled action through a discrete hidden Markov model. Action features have also been extracted through 3D Histograms of Oriented Displacements (HOD) [87], Accumulation of Motion Energy (AME) function aided with the EigenJoints method [61], and through a longest common sequence algorithm [88] to select high-discriminative power features from the relative motion trajectories of the skeleton.

### 3.4. Multi-Modal-Data-Based Techniques

The research results in [60,81,89] show that depth-based methods achieve better action-recognition performance than RGB-based methods. Therefore, some researchers have also tried a fusion of different modalities. Chaaroui *et al.* [90] have investigated the fusion of skeleton and depth data to overcome problems caused by occlusion and perspective changes in skeleton features. In addition to this, a sparse regression learning-based method to fuse depth and skeleton features has been proposed by Li *et al.* [62]. A multi-kernel-based learning method for describing actions has been proposed by Althloothi *et al.* [91] calculating spherical harmonics through depth data, and fused this with the spatial information of skeleton joints. Furthermore, RGB and depth-data fusion has also been attempted by some researchers. For example, Liu *et al.* [92] used generic algorithms, Jalal *et al.* [93] merged spatio-temporal features, and Ni *et al.* [94] introduced the multi-level fusion of RGB and depth data features. The main concern with multi-modal fusion is that it adds more computational complexity



**Figure 3.** Hierarchy of action-recognition techniques based on handcrafted features that use classical machine learning

to the action-recognition algorithm. Different challenges associated with multi-modal data-fusion techniques are discussed in Section 5.1.

#### 4. Deep-Learning-Based Techniques

Computer-vision researchers have directed considerable attention to the application of deep learning in action recognition. The classical machine-learning-based methods are based on handcrafted features, which are not robust. Deep-learning-based methods have been utilized due to their automated feature-learning from images. Researchers have extracted action features from RGB data, depth data, and skeleton sequences using deep-learning methods. Deep learning can directly obtain hierarchical features from different data modalities and provides a more effective solution. Accordingly, appearance and optical sequences can be used as inputs to deep networks. Besides aspects of appearance and motion information, deep-learning-based methods can also be applied using depth sequences and skeleton joint information. Wang *et al.* [95] have used convolution to learn action features from depth data. Wang *et al.* [96] combined motion and structure information in a depth sequence by pairing structured dynamic images at the body, part and joint levels through bidirectional rank pooling. Every pair is constructed from depth maps at each granularity level and serves as input to CNN. Song *et al.* [97] have proposed a model that uses different levels of attention in addition to a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) to learn discriminative skeleton joints. Ye *et al.* [98] have embedded temporal information with dense motion trajectories to learn actions. Yan *et al.* [99] have modeled relationships between graphs and joints by using graph-oriented CNN. Deep-learning-based feature learning has been shown to provide better performance than handcrafted feature extraction methods; however, there are still challenges concerning multi-modal data fusion. Deep-learning-based action recognition methods use different standalone as well as hybrid neural network architectures, which can be classified as Single-Stream, Two-Stream, LRCN, and Hybrid network-based architectures. The following sub-sections will summarize these architectural styles.

##### 4.1. Single Stream

A single-stream model is similar to the AlexNet [100] type of image-classification network. Single-stream architecture can take advantage of regularization through local filters, parameter-sharing at convolution layers, and local invariance-building neurons (max pooling). Such neural network architecture shifts the engineering focus from feature design strategies to network structure and

hyperparameter tuning strategies. Architectural details from AlexNet [100] can be used with different hyperparameter configurations. A single-stream architecture fuses information from all the frames in the softmax layer connected to the last fully connected layers with dense connections. Given an entire action video, the video-level prediction can be produced by forward propagation of each frame individually through the network and then averaging individual frame predictions over the duration of the video. However, single-stream architecture has been a foundation for other extended architectures. Some possible extensions to single-stream architecture have been explored by Baccouche *et al.* [101], Ji *et al.* [102] and Karpathy *et al.* [26].

#### 4.2. Two Stream

The two-stream model uses two disjointed CNNs containing spatial and temporal information, which are later fused together. The spatial network performs action recognition from single video frames, while the temporal network learns to recognize action from motion, i.e., dense optical flow. The idea behind this two-stream model relates to the fact that the human visual cortex contains two pathways for object and motion recognition, i.e., the ventral stream performs object recognition and the dorsal stream recognizes motion. Spatial-stream CNN is modeled similar to the single-frame model discussed earlier. Given an action video, each frame is individually passed through the spatial network where an action label is assigned to each frame. The Temporal-stream CNN is not the same as motion-aware CNN models (which use stacked single video frames as input). It takes stacked optical flow displacement fields between several consecutive frames as input to explicitly learn a temporal feature. There are three variations of optical flow-based inputs:

##### Optical Stacking

Input of temporal stream CNN is formed by stacking the dense optical flow of  $L$  consecutive frames. The optical flow at point  $(u, v)$  in frame  $t$  is a 2D displacement vector (i.e., horizontal and vertical displacement), which moves the point to the corresponding point in the next frame  $t + 1$ . Note that the horizontal and vertical components of the dense optical flow of a frame can be seen as image channels. Thus, the stacked dense optical flow of  $L$  consecutive frames forms input image  $2 \times L$  channels, which are then fed to the temporal stream CNN as input.

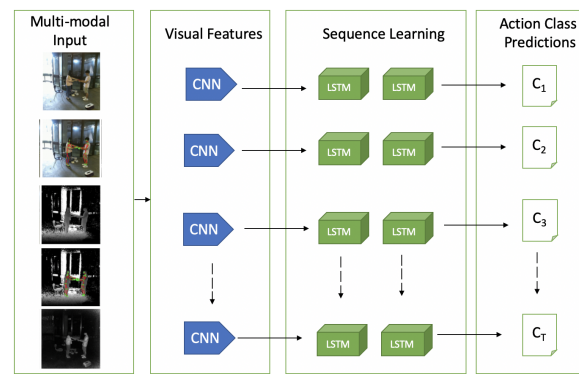
##### Trajectory Stacking

Unlike the optical stacking method, which samples displacement vectors at the same location in  $L$  consecutive frames, the trajectory-stacking method represents motion in an input image of  $2 \times L$  channels by sampling  $L$  2D points along motion trajectories [9].

##### Bi-directional Stacking

Both optical flow and trajectory stacking methods operate on forward optical flow. The bi-directional stacking method extends these methods by computing both forward and backward displacement of optical flow fields. More precisely, motion information is encoded in an input image of  $2 \times L$  channels stacking  $L/2$  forward optical flows between frames  $t$  and  $t + L/2$  and  $L/2$  backward optical flows between frames  $t - L/2$  and  $t$ .

The architectures of temporal and spatial CNN models are similar, except that the second normalization layer is removed from the temporal CNN to reduce memory consumption. Further, a class score fusion is added to the end of the models through late fusion. Different approaches can be used for class score fusion; however, experimental results show that training a linear SVM classifier on stacked L2-normalized softmax scores outperforms simple averaging. In two-stream models, the pioneering work of Simonyan and Zisserman [27] uses a single image and multi-optical flow sequence stack as input to the 2D CNN. Zhang *et al.* [103] have extended Simonyan and Zisserman's [27] work by using a motion vector instead of optical flow as an input to improve performance and comprehend



**Figure 4.** A possible architecture of LRCN with multi-modal input. Multi-modal Input visuals taken from [40]

real-time action recognition. Feichtenhofer *et al.* [104] have proposed an innovative approach involving moving the classification layer to the middle of the network for spatio-temporal information fusion, and this was shown to improve the accuracy. Wang *et al.* [105] have contributed to the input and training strategy of convolution and proposed Temporal Segment Network (TSN), improving the two-stream CNN. The notion of TSN was based on long-range temporal structural modelling. Later, Lan [106] and Zhou [107] enhanced the TSN. Carreira *et al.* [50] adopted the structure of Inception-v1 and inflated two-stream CNN to 3D CNN for action recognition. Zhu *et al.* [108] have expanded two-stream CNN to a 3D structure by drawing out the pooling operation.

#### 4.3. Long-term Recurrent Convolutional Network (LRCN)

LRCN uses CNN in co-ordination with an LSTM-based network. In LSTM-based deep-learning methods, where actions can be represented as feature changes between frames in the video. LSTM is widely used to improve action recognition techniques. Ng *et al.* [109] have presented a linear RNN for recognizing human action that connects the output of a CNN with an LSTM cell. A new architecture P3D ResNet has been proposed by Qiu *et al.* [110], which uniquely places all the variants of blocks in a different placement of ResNet. In skeletal data, to deal with noise, Liu *et al.* [111] extended the idea to analyze Spatio-temporal domains simultaneously by introducing an effective tree-structure based traversal framework. This framework uses a multimodal feature fusion strategy within LSTM unit and a gating mechanism to learn sequential data's reliability in long-term context representation. For mapping video frames with variable length inputs to variable length outputs, Donahue *et al.* [112] have proposed an LRCN. Unlike those methods which learn CNN filters based on a stack of a fixed number of input frames, LRCN [112] is not constrained to fixed-length input frames and thus can learn to recognize more complex action video. As illustrated in Fig. 4, in LRCN, individual video frames are first passed through CNN models with shared parameters and are then connected to a single-layer LSTM network. More precisely, the LRCN model combines a deep hierarchical visual-feature extractor, i.e., a CNN feature extractor, with an LSTM that can learn to recognize temporal variations in an end-to-end fashion.

#### 4.4. Hybrid Deep-Learning-Based Techniques for Action Recognition

Researchers have designed modern hybrid deep-learning-based techniques that have significantly improved outcomes. These hybrid models use the best of modalities and available network architectures to produce the best action recognition performance. Deep-learning-based competitive action recognition techniques which outperform other techniques on popular datasets are discussed below.

Abu-El-Haija *et al.* [48] pre-trained an LSTM network on a large-scale YT-8M dataset and discovered that pre-training on a large-scale data set generalizes well on datasets such as Sports-1M [26]

**Table 3.** List of popular action-recognition techniques with their key properties. Notations: MM: Multi-Modality; CA: Code Availability; ETD: Extra Training Data; TL: Transfer Learning; Y: Yes; N: No; mAP: Mean Average Precision; OF: Optical Flow; IF: Infrared; FV: Fisher Vectors; BoW: Bag of Words

Ref.	MM	CA	ETD	TL	Metric	Methods	Modality
[48]	N	Y	Y	Y	mAP	LSTM	RGB
[50]	N	Y	Y	Y	3-Fold Accuracy	i3D CNN	RGB
[51]	N	N	N	N	mAP	CNN	RGB
[33]	Y	Y	N	N	Accuracy	2D CNN	RGB+OF
[52]	Y	N	N	Y	Accuracy	3D CNN	RGB + OF
[43]	Y	Y	Y	Y	Accuracy	i3D CNN	RGB+Pose
[53]	N	Y	Y	Y	mAP	i3D CNN + Attention	RGB
[38]	N	Y	Y	Y	Accuracy	2D CNN	RGB
[54]	Y	N	Y	Y	mAP	CNN	RGB +OF
[113]	Y	Y	N	N	Accuracy	CNN	Skeleton
[41]	N	N	-	Y	Accuracy	i3D CNN	IF+OF+RGB
[55]	N	N	N	Y	Accuracy	3D CNN	RGB
[42]	N	N	N	N	Accuracy	GCN	Skeleton
[56]	N	N	Y	N	Accuracy	CNN	iDT/FV/BoW
[57]	N	Y	N	Y	Accuracy	CNN	RGB

and ActivityNet [29]. Jinyoung *et al.* [49] introduced an extensible hierarchical method for detecting generic interactive actions (by combining spatial relations with movements between two objects) and inherited actions determined through an ontology of rule-based methodology. This technique outperforms other techniques on the ActionNet-VE dataset [30].

Yuan *et al.* [51] have proposed the Multi-Granularity Generator (MGG) which produces temporal action proposals through two producers, namely the Segment Proposal Producer (SPP) and the Frame Actionness Producer (FAP). SPP generates a segment proposal from a coarse perspective, while FAP produces a finer actionness evaluation for each video frame. Both producers are combined to reflect two different granularities. MGG can be trained in end-to-end fashion and performs better than state-of-the-art methods on the THUMOS-14 dataset.

Zhaofan *et al.* [114] have provided a novel neural network architecture that uses spatio-temporal representation learning by Local and Global Diffusion (LGD) in parallel. This architecture is composed of LGD blocks, which update local and holistic features. In addition to this, a kernelized classifier is used for video recognition.

Lin *et al.* [33] have proposed an efficient and generic Temporal Shift Module (TSM) that claims the performance of 3D CNN while maintaining the complexity of 2D CNN. TSM facilitates information-sharing among neighboring frames by shifting a portion of the channels along the temporal dimension. TSM has been shown to achieve over state-of-the-art techniques on the Something-Something dataset.

Zhu *et al.* [43] have initiated the idea of an Action Machine which is a simple fast method extended from Inflated 3D CNN by adding a module of 2D CNN and pose estimation. Action Machine takes input that is cropped through person-bounding boxes, fusing predictions from RGB images and poses. Action Machine has been shown to achieve state-of-the-art results on NTU RGB-D datasets and competitive results on smaller datasets of Northwestern, MSR-DailyActivity3D, and UTD-MHAD [35,115].

Girdhar *et al.* [53] have proposed the Video Action Transformer Network (VATN) which uses transformer-style architecture that aggregates features from the spatio-temporal context. VATN uses an attention mechanism that learns to emphasize hands and faces. On the Atomic Visual Actions (AVA) dataset, VATN outperformed state-of-the-art methods by using raw RGB frames as input.

Hu *et al.* [116] have explored the modality-temporal mutual information to learn time-varying information and multi-modal features jointly. They introduced an action feature called a modality-temporal cube, which characterises RGB-D actions from a comprehensive perspective. Their

proposed framework uses deep bilinear blocks that pools input from both modality and temporal directions.

Sudhakaran *et al.* [38] have used spatial gating in the spatio-temporal decomposition of 3D kernels by implementing the concept of the Gate Shift Module (GSM). GSM is added to a 2D CNN that learns route features and combines them with less computational complexity and additional parameters overhead. This technique achieves state-of-the-art results on the Something-Something-v1 dataset and competitive results on other datasets.

Caetano *et al.* [113] have proposed the Skelemotion, which extends Spatio-Temporal Graph Convolutional Networks (ST-GCN) by introducing a Graph Vertex Feature Encoder (GVFE) and Dilated Hierarchical Temporal Convolutional Network (DH-TCN). GVFE learns vertex features by encoding raw skeleton features data, while DH-TCN captures both short-term and long-term dependencies. This architecture uses fewer layers and parameters and competes better with state-of-the-art methods on NTU RGB-D 60 and NTU RGB-D 120 datasets.

Korbar *et al.* [41] have introduced a clip-sampling scheme that selects salient temporal clips within a long video. This technique improves the state-of-the-art and reduces computational costs significantly.

Wang *et al.* [55] have proposed Channel Separated Convolutional Networks (CSNN) which demonstrates the benefits of factorizing 3D convolutions by separating spatio-temporal interactions and channel interactions. The latter is a form of regularization that improves accuracy and lowers computational costs.

Kim *et al.* [42] have introduced the Object-Related Human Action recognition through Graph Convolution Networks (OHA-GCN) which construct graphs using selective sampling of human and object poses. OHA-GCN late fuses class scores from human poses and object pose streams for action classification.

Wang *et al.* [56] have proposed a network that creates iDT descriptors and I3D optical flow features with CNNs, thus reviving classical handcrafted representations.

Liu *et al.* [57] have introduced a novel neural network called CPNet that learns evolving 2D fields with temporal consistency. CPNet achieved state-of-the-art results on both Jester [46] and Something-Something datasets [33,38].

Martin *et al.* [117] have introduced the novel approach for fine-grained categorization of driver behaviour. They focused on key challenges such as recognition of fine-grained behaviour inside the vehicle cabin; multi-modal activity recognition, focusing on diverse data streams; and a cross-view recognition benchmark. Adopting prominent methods for video and body pose-based action recognition to provide challenging benchmarks.

Munro and Damen [118] have exploited the correspondence of modalities as a self-supervised alignment approach in addition to adversarial alignment, which outperforms other unsupervised domain adaptation methods.

Table 3 summarizes the key properties of the best deep-learning-based action-recognition methods that have evolved over the last decade. It can be observed that most of the techniques are not applied on multi-modal datasets. Accuracy is the preferred metric for action recognition, where sometimes these techniques benefit from the use of extra training data.

In Human Action Recognition, the implementation and execution of deep-learning-based methods can often be time-consuming. Experimental platforms provide abstraction, customization, community, and advanced hardware-level support. This is important for the development of robust and flexible deep-learning-based action recognition systems. Some platforms are intuitive and highly abstract, but such abstractions or wrappers can make it difficult to debug or apply explicit changes to algorithms at low levels. As performance demand relies on high-end hardware, multi-GPU support is a must when experimenting with big-data-related problems.

## 5. Challenges and Future Research Directions

Action recognition remains challenging due to background clutter, partial occlusion, viewpoint, lighting changes, execution rate, and biometric variation. Challenges with data-fusion, and action recognition, followed by future research directions, are discussed below.

### 5.1. Challenges in Data-Fusion and Action-Recognition Techniques

Data-fusion in deep-learning architectures uses approaches such as early, slow, late fusion, and other variants. However, an innovative strategy for combining or augmenting of different modalities at earlier or any later phase of resolution can lead to better data-fusion solutions. Action datasets with different resolutions are an inherent challenge in data fusion because each modality has a very different temporal and spatial resolution. Practically, individual datasets contain incompatible numbers of data samples, which leads to data-size incompatibility. Alignment of modalities to a standard coordinate system for maximizing mutual information sharing is an acute challenge in data fusion.

Inherently, the information conveyed through each modality has different physical properties, which can be vital for better action learning. Other challenging factors include noise, spatial distortions, varying contrast, and arbitrary subject locations in image sequences. Negligible errors produced by multi-modal sensors are often abstracted as noise, which is unavoidable. Balancing noise with other modalities also causes problems in data fusion. Most data-fusion techniques ignore the noise, but ignoring the noise from datasets collected through different sensors may lead to bias. Further, distinct data modalities confront contradictions, and data inconsistencies may occur. An open challenge is to infer a proper compromise; however, identifying these conflicts, contradictions, and inconsistencies is a fundamental challenge. Similarly, sensors may produce spurious data due to environmental or sensor failure issues, which may lead to false inferences based on biased estimations. In data fusion, a challenge may arise in predicting and modeling spurious events.

Multi-modal action-recognition techniques vary across different tasks and need to address various challenges in terms of required training time and memory management. For example, The use of neural blocks (10, in most cases) causes a significant increase in the number of parameters and, consequently, the memory requirement and computational complexity for techniques that use ST-GCN in addition to their handcrafted classification modules [119]. Additional problems arise depending on the modality under consideration. For skeleton datasets, the use of basic skeleton features (joint coordinates or bone lengths) is common for constructing spatial-temporal graphs. However, offering a high-level description of the human body structure may affect discriminative power for action-recognition.

### 5.2. Future Research Directions

The discussion and insights drawn from the challenges in different approaches allow us to present several future research directions to develop methods in action recognition. The following research directions may advance the domain:

**Combination of classical-machine learning and deep learning-based methods.** Classical machine learning approaches have benefited action recognition through redundant and favourable feature extraction. Deep learning-based methods provide autonomous feature engineering and have produced better recognition systems. Designing effective action recognition systems by adding the power of classical machine learning with advanced deep learning-based techniques has some attraction for researchers. For example, Gao *et al.* [120] proposed a fusion logic of classical machine learning and deep learning-based methods to achieve better performance than single CNN-based pedestrian detector, and it is likely to emerge as an active research area.

**Assessment in practical scenarios.** Most of the uni-modal and multi-modal datasets have been collected in constraint environments. There remains a significant gap between the collected datasets during the last few years and the practical scenario due to insufficient categories, occlusion cases, constrained environment settings, samples, and limited distance variations. Due to these limitations,

collected datasets may not substitute the need for outdoor practical scenario-based datasets. Collection and generalization of algorithms over realistic scenarios should gain the attention of researchers.

**Self-learning.** Learning labels about individual samples is often overlapping and causes inefficient intra-class similarity. Self-learning action recognition systems can learn from non-labelled training data without any human intervention. Recent advances in deep learning, such as Generative Adversarial Networks (GAN), may improve action recognition systems' self-learning capability. GAN-based action recognition techniques would be an compelling research direction.

**Interpretation of online human actions.** Action recognition algorithms focus on well-trimmed segmented data splits. While in an online action recognition system, which aims to observe many mechanisms such as facial expression, visual focus, view angles etc. instantly from a video stream. Interpretation of such human behaviour components in online scenarios is an essential step toward more practical and intelligent recognition systems.

**Multi-modal fusion.** Multi-modal data provides richer information than uni-modal data. Still, most methods fuse different modalities as separate channels and combine them at a later classification stage without exploiting their corresponding properties in a parallel fashion. Effective use of deep networks for parallel integration of complementary properties from different modalities would a potential research area. Use of multi-modal information also helpful in reducing noise from uni-modal data. Therefore, integrating multi-modal information and incorporating contextual information from the surrounding environment is a way forward for future research. Different fusion schemes are used in various methods for action classification. Thus, future research may devote more attention from researchers to compare these fusion schemes and find the best fusion strategy for action recognition.

## 6. Conclusion

The vision-based recognition of human actions is an important research field in the integrative computer vision and multimedia analytics ecosystem. This review has thoroughly compared and summarized the landscape of vision-based multi-modal action data capture sensors. We gave a brief review of existing commonly used datasets and highlighted key research that has mainly focused on multi-modal datasets. We also reviewed the latest action detection techniques that use deep learning in general. We discussed the techniques that have been used over the past decade and divided them into different perspectives. We then presented various available experimental options, along with their characteristics, strengths and weaknesses, for action recognition researchers.

The results of this paper show that with the availability of low-cost, multi-function sensors, the effects of multi-modal action detection can be extended to wider application areas. It is evident that deep learning architectures, especially CNN and LSTM-based methods, have been shown to produce significant results. However, there is a lack of availability of large data sets in different domains. Attention has turned more to RGB, optical flow, and skeletal modalities, so other promising modalities such as depth and IR have not been adequately explored. The challenges are evident with multi-modal action sensors, data sets, detection and fusion techniques. Significant efforts are required to address these challenges.

**Author Contributions:** conceptualization, M.S. and D.C.; methodology, M.S.; software, M. S.; validation, D. C.; formal analysis, M. S.; investigation, M. S.; data curation, M. S.; writing—original draft preparation, M. S.; writing—review and editing, D. C.; visualization, M. S.; supervision, D. C.; project administration, D. C.; funding acquisition, D. C.

**Funding:** This research was funded by a joint research project grant [No.5-1/HRD/UESTPI(Batch-VI)/7108/2018/HEC] of the Higher Education Commission (HEC) Pakistan and Edith Cowan University (ECU) Australia.

**Acknowledgments:** The authors would like to thank anonymous reviewers for their careful reading and valuable remarks, which have greatly helped extend the scope of this paper. All statements of fact, opinion, or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of the sponsors.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

AME	Accumulation of Motion Energy
AVA	Atomic Visual Actions
CSNN	Channel Separated Convolutional Networks
CNN	Convolutional Neural Network
DH-TCN	Dilated Hierarchical Temporal Convolutional Network
FAP	Frame Actionness Producer
GPU	Graphics Processing Unit
GVFE	Graph Vertex Feature Encoder
HOD	Histograms of Oriented Displacements
iDT	improved Dense Trajectories
IR	Infrared
HOG	Histogram Of Gradients
LGD	Local and Global Diffusion
LRCN	Long-term Recurrent Convolutional Network
LSTM	Long-Short Term Memory
mAP	mean Average Precision
MEI	Motion Energy Image
MGG	Multi-Granularity Generator
MHI	Motion History Image
OHA-GCN	Object-Related Human Action recognition through Graph Convolution Networks
RGB	Red Green Blue
RNN	Recurrent Neural Network
SPP	Segment Proposal Producer
STIP	Spatio Temporal Interesting Points
ST-GCN	Spatio Temporal Graph Convolutional Networks
SVM	Scalar Vector Machines
TSM	Temporal Shift Module
VATN	Video Action Transformer Network

## References

1. Inc, A.C. Xtion PRO LIVE | 3D Sensor | ASUS USA. [https://www.asus.com/us/3D-Sensor/Xtion\\_PRO\\_LIVE/](https://www.asus.com/us/3D-Sensor/Xtion_PRO_LIVE/), 2020. Accessed: 07-01-2021.
2. Yang, L.; Zhang, L.; Dong, H.; Alelaiwi, A.; Saddik, A.E. Evaluating and Improving the Depth Accuracy of Kinect for Windows v2. *IEEE Sensors* **2015**, *15*, 4275–4285.
3. Carfagni, M.; Furferi, R.; Governi, L.; Santarelli, C.; Servi, M.; Uccheddu, F.; Volpe, Y. Metrological and Critical Characterization of the Intel D415 Stereo Depth Camera. *Sensors* **2019**, *19*, 489.
4. Herath, S.; Harandi, M.; Porikli, F. Going Deeper into Action Recognition: A Survey. *Image Vis. Comput.* **2017**, *60*, 4–21.
5. Aggarwal, J.; Cai, Q. Human Motion Analysis: A Review. *Comput. Vis. Image Underst.* **1999**, *73*, 428–440.
6. Guo, G.; Lai, A. A Survey on Still-Image-based Human Action Recognition. *Pattern Recognit.* **2014**, *47*, 3343–3361.
7. Poppe, R. A Survey on Vision-based Human Action Recognition. *Image Vis. Comput.* **2010**, *28*, 976–990.
8. Turaga, P.; Chellappa, R.; Subrahmanian, V.S.; Udrea, O. Machine Recognition of Human Activities: A Survey. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1473–1488.
9. Wang, H.; Kläser, A.; Schmid, C.; Cheng-Lin, L. Action Recognition by Dense Trajectories. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Providence, USA, 2011; pp. 3169–3176.

10. Zhu, G.; Zhang, L.; Mei, L.; Jie Shao.; Juan Song.; Peiyi Shen. Large-scale Isolated Gesture Recognition using Pyramidal 3D Convolutional Networks. 23rd Int. Conf. on Pattern Recognit.; IEEE: Cancun, Mexico, 2016; pp. 19–24.
11. Asadi-Aghbolaghi, M.; Clapés, A.; Bellantonio, M.; Escalante, H.J.; Ponce-López, V.; Baró, X.; Guyon, I.; Kasaei, S.; Escalera, S. A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences. In Proceedings Int. Conf. on Automatic Face Gesture Recognition; IEEE: Washington, USA, 2017; pp. 476–483.
12. Prince, S. *Computer Vision: Models, Learning, and Inference*, 1st ed.; Cambridge University Press: USA, 2012.
13. Szeliski, R. *Computer Vision: Algorithms and Applications*, 1st ed.; Springer: UK, 2010.
14. Wang, P.; Li, W.; Ogunbona, P.; Wan, J.; Escalera, S. RGB-D-based Human Motion Recognition with Deep Learning: A Survey. *Comput. Vis. Image Underst.* **2018**, *171*, 118–139.
15. Aggarwal, J.; Xia, L. Human Activity Recognition from 3D Data: A Review. *Pattern Recognit. Lett.* **2014**, *48*, 70–80.
16. Chen, L.; Wei, H.; Ferryman, J. A Survey of Human Motion Analysis using Depth Imagery. *Pattern Recognit. Lett.* **2013**, *34*, 1995–2006.
17. Han, F.; Reily, B.; Hoff, W.; Zhang, H. Space-time Representation of People based on 3D Skeletal Data: A Review. *J. Vis. Commun. Image Represent.* **2017**, *158*, 85–105.
18. Zhang, J.; Li, W.; Ogunbona, P.O.; Wang, P.; Tang, C. RGB-D-based Action Recognition Datasets: A Survey. *Pattern Recognit.* **2016**, *60*, 86–105.
19. Ye, M.; Zhang, Q.; Wang, L.; Zhu, J.; Yang, R.; Gall, J. A Survey on Human Motion Analysis from Depth Data. In: Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications, Lecture Notes in Computer Science; Springer: Berlin, Germany, 2013; Vol. 8200, pp. 149–187.
20. Zhu, F.; Shao, L.; Xie, J.; Fang, Y. From Handcrafted to Learned Representations for Human Action Recognition: A Survey. *Image and Vis. Comput.* **2016**, *55*, 42–52.
21. Zhang, H.B.; Zhang, Y.X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.X.; Chen, D.S. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. *Sensors* **2019**, *19*, 1005.
22. Chen, C.; Jafari, R.; Kehtarnavaz, N. A Survey of Depth and Inertial Sensor Fusion for Human Action Recognition. *Multimedia Tools and Applications* **2017**, *76*, 4405–4425.
23. Zhang, Z.; Ma, X.; Song, R.; Rong, X.; Tian, X.; Tian, G.; Li, Y. Deep Learning-based Human Action Recognition: A Survey. Chinese Automation Congress (CAC); IEEE: Jinan, China, 2017; pp. 3780–3785.
24. Minh Dang, L.; Min, K.; Wang, H.; Jalil Piran, M.; Hee Lee, C.; Moon, H. Sensor-based and Vision-based Human Activity Recognition: A Comprehensive Survey. *Pattern Recognit.* **2020**, *108*, 107561/1–24.
25. Jiang, H.; Li, Y.; Song, S.; Liu, J. Rethinking Fusion Baselines for Multi-modal Human Action Recognition. In Proceedings of 19th Pacific-Rim Conference on Multimedia, Advances in Multimedia Information Processing; Springer: Hefei, China, 2018; pp. 178–187.
26. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale Video Classification With Convolutional Neural Networks. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Columbus, USA, 2014; pp. 1725–1732.
27. Simonyan, K.; Zisserman, A. Two-stream Convolutional Networks for Action Recognition in Videos. In Proceedings of The 27th Int. Conf. on Neural Information Processing Systems; MIT Press: Montreal, Canada, 2014; pp. 568–576.
28. Oh, S.; Hoogs, A.; Perera, A.; Cuntoor, N.; Chen, C.; Lee, J.T.; Mukherjee, S.; Aggarwal, J.K.; Lee, H.; Davis, L.; Swears, E.; Wang, X.; Ji, Q.; Reddy, K.; Shah, M.; Vondrick, C.; Pirsiavash, H.; Ramanan, D.; Yuen, J.; Torralba, A.; Song, B.; Fong, A.; Roy-Chowdhury, A.; Desai, M. A Large-scale Benchmark Dataset for Event Recognition in Surveillance Video. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Providence, USA, 2011; pp. 3153–3160.
29. Fabian Caba Heilbron, Victor Escorcia, B.G.; Nieves, J.C. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Boston, USA, 2015; pp. 961–970.
30. Moon, J.; Kwon, Y.; Kang, K.; Park, J. ActionNet-VE Dataset: A Dataset for Describing Visual Events by Extending VIRAT Ground 2.0. In Proceedings of The 8th Int. Conf. on Signal Processing, Image Processing and Pattern Recognit.; IEEE: Jeju, South Korea, 2015; pp. 1–4.

31. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, 2012, [[arXiv:cs.CV/1212.0402](https://arxiv.org/abs/cs.CV/1212.0402)].
32. Jiang, Y.G.; Liu, J.; Roshan Zamir, A.; Toderici, G.; Laptev, I.; Shah, M.; Sukthankar, R. THUMOS Challenge: Action Recognition with a Large Number of Classes. <http://csrcv.ucf.edu/THUMOS14/>, 2014.
33. Lin, J.; Gan, C.; Han, S. TSM: Temporal Shift Module for Efficient Video Understanding. In Proceedings of The Int. Conf. on Comput. Vis. (ICCV); IEEE: Seoul, South Korea, 2019; pp. 7083–7093.
34. Ohn-Bar, E.; Trivedi, M.M. Hand Gesture Recognition in Real Time for Automotive Interfaces: A Multimodal Vision-Based Approach and Evaluations. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 2368–2377.
35. Chen, C.; Jafari, R.; Kehtarnavaz, N. UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing A Depth Camera and A Wearable Inertial Sensor. In Proceedings of The Int. Conf. on Image Processing; IEEE: Quebec City, Canada, 2015; pp. 168–172.
36. Zhang, Y.; Cao, C.; Cheng, J.; Lu, H. EgoGesture: A New Dataset and Benchmark for Egocentric Hand Gesture Recognition. *IEEE Trans. Multimedia* **2018**, *20*, 1038–1050.
37. Gu, C.; Sun, C.; Ross, D.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; Schmid, C.; Malik, J. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE Computer Society: Los Alamitos, CA, USA, 2018; pp. 6047–6056.
38. Sudhakaran, S.; Escalera, S.; Lanz, O. Gate-Shift Networks for Video Action Recognition. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE Computer Society: Los Alamitos, USA, 2020; pp. 1099–1108.
39. Sigurdsson, G.A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; Gupta, A. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In Proceedings of European Conf. Comput. Vis. (ECCV); Springer: Amsterdam, Netherlands, 2016; pp. 510–526.
40. Liu, J.; Shahroudy, A.; Perez, M.L.; Wang, G.; Duan, L.Y.; Kot Chichung, A. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2019**, pp. 2684 – 2701.
41. Korbar, B.; Tran, D.; Torresani, L. SCSampler: Sampling Salient Clips from Video for Efficient Action Recognition. In Proceedings of The Int. Conf. on Comput. Vis. (ICCV); IEEE: Seoul, South Korea, 2019; pp. 6231–6241.
42. Kim, S.; Yun, K.; Park, J.; Choi, J. Skeleton-Based Action Recognition of People Handling Objects. In Proceedings of The IEEE Winter Conf. on Applications of Comput. Vis. (WCACV); IEEE: Hawaii, USA, 2019; pp. 61–70.
43. Zhu, J.; Zou, W.; Xu, L.; Hu, Y.; Zhu, Z.; Chang, M.; Huang, J.; Huang, G.; Du, D. Action Machine: Rethinking Action Recognition in Trimmed Videos, 2018, [[arXiv:cs.CV/1812.05770](https://arxiv.org/abs/cs.CV/1812.05770)].
44. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A Large Video Database for Human Motion Recognition. In Proceedings of The Int. Conf. on Comput. Vis. (ICCV); IEEE: Barcelona, Spain, 2011; pp. 2556–2563.
45. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE Computer Society: Los Alamitos, USA, 2016; pp. 1010–1019.
46. Materzynska, J.; Berger, G.; Bax, I.; Memisevic, R. The Jester Dataset: A Large-Scale Video Dataset of Human Gestures. In Proceedings of The Int. Conf. on Comp. Vis. Workshops; IEEE: Seoul, South Korea, 2019.
47. Wang, X.; Ji, Q. A Hierarchical Context Model for Event Recognition in Surveillance Video. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Columbus, USA, 2014; pp. 2561–2568.
48. Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; Vijayanarasimhan, S. YouTube-8M: A Large-Scale Video Classification Benchmark, 2016, [[arXiv:cs.CV/1609.08675](https://arxiv.org/abs/cs.CV/1609.08675)].
49. Moon, J.; Jin, J.; Kwon, Y.; Kang, K.; Park, J.; Park, K. Extensible Hierarchical Method of Detecting Interactive Actions for Video Understanding. *ETRI J.* **2017**, *39*, 502–513.
50. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Honolulu, USA, 2017; pp. 4724–4733.

51. Liu, Y.; Ma, L.; Zhang, Y.; Liu, W.; Chang, S. Multi-Granularity Generator for Temporal Action Proposal. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Long Beach, USA, 2019; pp. 3604–3613.
52. Abavisani, M.; Joze, H.; Patel, V. Improving The Performance of Unimodal Dynamic Hand-Gesture Recognition With Multimodal Training. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Long Beach, USA, 2019; pp. 1165–1174.
53. Girdhar, R.; João Carreira, J.; Doersch, C.; Zisserman, A. Video Action Transformer Network. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Long Beach, USA, 2019; pp. 244–253.
54. Ryoo, M.S.; Piergiovanni, A.; Tan, M.; Angelova, A. AssembleNet: Searching for Multi-Stream Neural Connectivity in Video Architectures, 2020, [[arXiv:cs.CV/1905.13209](https://arxiv.org/abs/1905.13209)].
55. Tran, D.; Wang, H.; Torresani, L.; Feiszli, M. Video Classification with Channel-separated Convolutional Networks. In Proceedings of The Int. Conf. on Comput. Vis. (ICCV); IEEE: Seoul, South Korea, 2019; pp. 5552–5561.
56. Wang, L.; Koniusz, P.; Huynh, D.Q. Hallucinating iDT Descriptors and i3D Optical Flow Features for Action Recognition with CNNs. In Proceedings of The Int. Conf. on Comput. Vis. (ICCV); IEEE: Seoul, South Korea, 2019; pp. 8698–8708.
57. Liu, X.; Lee, J.; Jin, H. Learning Video Representations From Correspondence Proposals. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Long Beach, USA, 2019; pp. 4273–4281.
58. Das, S.; Sharma, S.; Dai, R.; Bremond, F.; Thonnat, M. VPN: Learning Video-Pose Embedding for Activities of Daily Living, 2020, [[arXiv:cs.CV/2007.03056](https://arxiv.org/abs/2007.03056)].
59. Oreifej, O.; Liu, Z. HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Portland, USA, 2013; pp. 716–723.
60. Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time Human Action Recognition Based on Depth Motion Maps. *J. Real Time Image Process.* **2016**, *12*, 155–163.
61. Yang, X.; Tian, Y. Effective 3D Action Recognition using EigenJoints. *J. Vis. Commun. Image Represent.* **2014**, *25*, 2–11.
62. Li, M.; Leung, H.; Shum, H.P. Human Action Recognition via Skeletal and Depth based Feature Fusion. In Proceedings of The 9th Int. Conf. on Motion in Games; ACM: Burlingame, USA, 2016; pp. 123–132.
63. Nguyen, T.V.; Song, Z.; Yan, S. STAP: Spatial-Temporal Attention-Aware Pooling for Action Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *25*, 77–86.
64. Zhang, H.B.; Lei, Q.; Zhong, B.N.; Du, J.X.; Peng, J.; Hsiao, T.C.; Chen, D.S. Multi-surface Analysis for Human Action Recognition in Video. *SpringerPlus* **2016**, *5*, 1226.
65. Dawn, D.D.; Shaikh, S.H. A Comprehensive Survey of Human Action Recognition with Spatio-temporal Interest Point (STIP) Detector. *Vis. Comput.* **2016**, *32*, 289–306.
66. Zhang, Z.; Liu, S.; Liu, S.; Han, L.; Shao, Y.; Zhou, W. Human Action Recognition using Salient Region Detection in Complex Scenes. In Proceedings of 3rd Int. Conf. on Communications, Signal Processing, and Systems; Springer: Hohot, China, 2015; pp. 565–572.
67. Burghouts, G.J.; Schutte, K.; ten Hove, R.J.; van den Broek, S.P.; Baan, J.; Rajadell, O.; van Huis, J.R.; van Rest, J.; Hanckmann, P.; Bouma, H.; Sanroma, G.; Evans, M.; Ferryman, J. Instantaneous Threat Detection Based on a Semantic Representation of Activities, Zones and Trajectories. *Signal Image Video Process.* **2014**, *8*, 191–200.
68. Wang, H.; Schmid, C. Action Recognition with Improved Trajectories. In Proceedings of The Int. Conf. on Comput. Vis. (ICCV); IEEE: Sydney, Australia, 2013; pp. 3551–3558.
69. Bobick, A.F.; Davis, J.W. The Recognition of Human Movement using Temporal Templates. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2001**, *23*, 257–267.
70. Zhang, Z.; Hu, Y.; Chan, S.; Chia, L.T. Motion Context: A New Representation for Human Action Recognition. In Proceedings of The European Conf. on Comput. Vis. (ECCV); Springer: Marseille, France, 2008; pp. 817–829.
71. Klaeser, A.; Marszalek, M.; Schmid, C. A Spatio-Temporal Descriptor Based on 3D-Gradients. In Proceedings of The British Machine Vision Conference; BMVA Press: Leeds, UK, 2008; pp. 99.1–99.10.

72. Somasundaram, G.; Cherian, A.; Morellas, V.; Papanikolopoulos, N. Action Recognition using Global Spatio-temporal Features derived from Sparse Representations. *Comput. Vis. Image Underst.* **2014**, *123*, 1–13.
73. Laptev, I. On space-time interest points. *Int. J. Comput. Vis.* **2005**, *64*, 107–123.
74. Chakraborty, B.; Holte, M.B.; Moeslund, T.B.; González, J. Selective spatio-temporal interest points. *Comput. Vis. Image Underst.* **2012**, *116*, 396 – 410.
75. Peng, X.; Wang, L.; Wang, X.; Qiao, Y. Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice. *Comput. Vis. Image Underst.* **2016**, *150*, 109–125.
76. Nazir, S.; Yousaf, M.H.; Velastin, S.A. Evaluating a Bag-of-visual Features Approach using Spatio-temporal Features for Action Recognition. *Comput. Electr. Eng.* **2018**, *72*, 660–669.
77. Gaidon, A.; Harchaoui, Z.; Schmid, C. Activity Representation with Motion Hierarchies. *Int. J. Comput. Vis.* **2014**, *107*, 219–238.
78. Wang, H.; Oneata, D.; Verbeek, J.; Schmid, C. A Robust and Efficient Video Representation for Action Recognition. *Int. J. Comput. Vis.* **2016**, *119*, 219–238.
79. Peng, X.; Zou, C.; Qiao, Y.; Peng, Q. Action Recognition with Stacked Fisher Vectors. In Proceedings of The European Conf. on Comput. Vis. (ECCV); Springer: Zurich, Switzerland, 2014; pp. 581–595.
80. Yang, X.; Tian, Y. Super Normal Vector for Activity Recognition using Depth Sequences. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Columbus, USA, 2014; pp. 804–811.
81. Rahmani, H.; Mahmood, A.; Huynh, D.Q.; Mian, A. Real Time Action Recognition using Histograms of Depth Gradients and Random Decision Forests. In Proceedings of The IEEE Winter Conf. on Applications of Comput. Vis. (WCACV); IEEE: SteamBoats Springs, USA, 2014; pp. 626–633.
82. Yang, X.; Zhang, C.; Tian, Y. Recognizing Actions using Depth Motion Maps-based Histograms of Oriented Gradients. In Proceedings of The 20th ACM Int. Conf. on Multimedia; ACM: Nara, Japan, 2012; pp. 1057–1060.
83. Chen, C.; Jafari, R.; Kehtarnavaz, N. Action Recognition from Depth Sequences using Depth Motion Maps-based Local Binary Patterns. In Proceedings of The IEEE Winter Conf. on Applications of Comput. Vis. (WCACV); IEEE: Waikola, USA, 2015; pp. 1092–1099.
84. Chen, W.; Guo, G. TriViews: A General Framework to use 3D Depth Data Effectively for Action Recognition. *J. Vis. Commun. Image Represent.* **2015**, *26*, 182–191.
85. Miao, J.; Jia, X.; Mathew, R.; Xu, X.; Taubman, D.; Qing, C. Efficient Action Recognition from Compressed Depth Maps. In Proceedings of The IEEE Int. Conf. on Image Processing; IEEE: Phoenix, USA, 2016; pp. 16–20.
86. Xia, L.; Chen, C.; Aggarwal, J. View Invariant Human Action Recognition using Histograms of 3D Joints. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops; IEEE: Providence, USA, 2012; pp. 20–27.
87. Gowayyed, M.A.; Torki, M.; Hussein, M.E.; El-Saban, M. Histogram of Oriented Displacements (HOD): Describing Trajectories of Human Joints for Action Recognition. In Proceedings of The 23rd Int. Joint Conf. on Artificial Intelligence; ACM: Beijing, China, 2013.
88. Pazhoumand-Dar, H.; Lam, C.P.; Masek, M. Joint Movement Similarities for Robust 3D Action Recognition using Skeletal Data. *J. Vis. Commun. Image Represent.* **2015**, *30*, 10–21.
89. Papadopoulos, G.T.; Axenopoulos, A.; Daras, P. Real-time Skeleton-tracking-based Human Action Recognition using Kinect Data. In Proceedings of The Int. Conf. on Multimedia Modeling; Springer: Dublin, Ireland, 2014; pp. 473–483.
90. Chaaraoui, A.; Padilla-Lopez, J.; Flórez-Revuelta, F. Fusion of Skeletal and Silhouette-based Features for Human Action Recognition with RGB-D Devices. In Proceedings of The IEEE Int. Conf. on Comput. Vis. (ICCV) Workshops; IEEE: Sydney, Australia, 2013; pp. 91–97.
91. Althloothi, S.; Mahoor, M.H.; Zhang, X.; Voyles, R.M. Human Activity Recognition using Multi-features and Multiple Kernel Learning. *Pattern Recognit.* **2014**, *47*, 1800–1812.
92. Liu, L.; Shao, L. Learning Discriminative Representations from RGB-D Video Data. In Proceedings of The Int. Joint Conf. on Artificial Intelligence; AAAI Press: Beijing, China, 2013; p. 1493–1500.
93. Jalal, A.; Kim, Y.H.; Kim, Y.J.; Kamal, S.; Kim, D. Robust Human Activity Recognition from Depth Video using Spatiotemporal Multi-fused Features. *Pattern Recognit.* **2017**, *61*, 295–308.

94. Ni, B.; Pei, Y.; Moulin, P.; Yan, S. Multilevel Depth and Image Fusion for Human Activity Detection. *IEEE Trans. Syst. Man Cybern.* **2013**, *43*, 1383–1394.
95. Wang, P.; Li, W.; Gao, Z.; Zhang, J.; Tang, C.; Ogunbona, P. Deep Convolutional Neural Networks for Action Recognition Using Depth Map Sequences, 2015, [[arXiv:cs.CV/1501.04686](https://arxiv.org/abs/1501.04686)].
96. Wang, P.; Wang, S.; Gao, Z.; Hou, Y.; Li, W. Structured Images for RGB-D Action Recognition. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops; IEEE: Venice, Italy, 2017; pp. 1005–1014.
97. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. Spatio-Temporal Attention-Based LSTM Networks for 3D Action Recognition and Detection. *IEEE Trans. Image Process. (TIP)* **2018**, *27*, 3459–3471.
98. Ye, Y.; Tian, Y. Embedding Sequential Information into Spatiotemporal Features for Action Recognition. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops; IEEE: Las Vegas, USA, 2016; pp. 1110–1118.
99. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of The 32nd Association for The Advancement of Artificial Intelligence (AAAI); AAAI Press: New Orleans, USA, 2018; pp. 8561–8568.
100. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *In Proceedings of Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Lake Tahoe, USA, 2012; pp. 1097–1105.
101. Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; Baskurt, A. Sequential Deep Learning for Human Action Recognition. In Proceedings of Human Behavior Understanding: 2nd International Workshop; Springer: Berlin, Germany, 2011; pp. 29–39.
102. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2013**, *35*, 221–231.
103. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-Time Action Recognition With Deeply Transferred Motion Vector CNNs. *IEEE Trans. Image Process. (TIP)* **2018**, *27*, 2326–2339.
104. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional Two-Stream Network Fusion for Video Action Recognition. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Las Vegas, USA, 2016.
105. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Gool, L.V. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In Proceedings of The European Conf. on Comput. Vis. (ECCV); Springer: Amsterdam, Netherlands, 2016; pp. 20–36.
106. Lan, Z.; Zhu, Y.; Hauptmann, A.G.; Newsam, S. Deep Local Video Feature for Action Recognition. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Honolulu, USA, 2017; pp. 1219–1225.
107. Zhou, B.; Andonian, A.; Torralba, A. Temporal Relational Reasoning in Videos. In Proceedings of The European Conf. on Comput. Vis. (ECCV); Springer: Munich, Germany, 2018; Vol. 11205, pp. 831–846.
108. Zhu, Y.; Lan, Z.; Newsam, S.; Hauptmann, A. Hidden Two-Stream Convolutional Networks for Action Recognition. In Proceedings of Asian Conf. on Comput. Vis.; Springer: Perth, Australia, 2018; pp. 363–378.
109. Ng, J.Y.H.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond Short Snippets: Deep Networks for Video Classification. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Boston, USA, 2015; pp. 4694–4702.
110. Qiu, Z.; Yao, T.; Mei, T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In Proceedings of The Int. Conf. on Comput. Vis. (ICCV); IEEE: Venice, Italy, 2017; pp. 5533–5541.
111. Liu, J.; Shahroudy, A.; Xu, D.; Kot, A.C.; Wang, G. Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2018**, *40*, 3007–3021.
112. Donahue, J.; Hendricks, L.A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2017**, *39*, 677–691.
113. Caetano, C.; Sena de Souza, J.; Santos, J.; Schwartz, W. SkeleMotion: A New Representation of Skeleton Joint Sequences Based on Motion Information for 3D Action Recognition. In Proceedings of The Int. Conf. on Advanced Video and Signal-based Surveillance; IEEE: Taipei, Taiwan, 2019; pp. 1–8.

114. Qiu, Z.; Yao, T.; Ngo, C.W.; Tian, X.; Mei, T. Learning Spatio-Temporal Representation With Local and Global Diffusion. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Long Beach, USA, 2019; pp. 12056–12065.
115. Wang, J.; Nie, X.; Xia, Y.; Wu, Y.; Zhu, S.C. Cross-view Action Modeling, Learning and Recognition. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Columbus, OH, USA, 2014; pp. 2649–2656.
116. Hu, J.F.; Zheng, W.S.; Pan, J.; Lai, J.; Zhang, J. Deep Bilinear Learning for RGB-D Action Recognition. In Proceedings of The European Conf. Comput. Vis. (ECCV); Springer: Munich, Germany, 2018; pp. 346–362.
117. Martin, M.; Roitberg, A.; Haurilet, M.; Horne, M.; Reiß, S.; Voit, M.; Stiefelhagen, R. Drive & Act: A Multi-Modal Dataset for Fine-Grained Driver Behavior Recognition in Autonomous Vehicles. In Proceedings of Int. Conf. on Comput. Vis. (ICCV); IEEE: Seoul, South Korea, 2019; pp. 2801–2810.
118. Munro, J.; Damen, D. Multi-modal Domain Adaptation for Fine-grained Action Recognition. In Proceedings of The IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR); IEEE: Seattle, USA, 2020.
119. Papadopoulos, K.; Ghorbel, E.; Aouada, D.; Ottersten, B. Vertex Feature Encoding and Hierarchical Temporal Modeling in a Spatial-Temporal Graph Convolutional Network for Action Recognition, 2019, [[arXiv:cs.CV/1912.09745](https://arxiv.org/abs/cs.CV/1912.09745)].
120. Gao, F.; Wang, C.; Li, C. A Combined Object Detection Method With Application to Pedestrian Detection. *IEEE Access* **2020**, *8*, 194457–194465.

© 2021 by the authors. Submitted to *Journal Not Specified* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).