applied sciences

MDPI

Review

A Comprehensive Review on Handcrafted and Learning-Based Action Representation Approaches for Human Activity Recognition

Allah Bux Sargano [1,2,*], Plamen Angelov [1] and Zulfiqar Habib [2]

[1] School of Computing and Communications Infolab21, Lancaster University, Lancaster LA1 4WA, UK; p.angelov@lancaster.ac.uk

[2] Department of Computer Science, COMSATS Institute of Information Technology, Lahore 54000, Pakistan; drzhabib@ciitlahore.edu.pk

* Correspondence: a.bux@lancaster.ac.uk; Tel.: +44-152-451-0525

Academic Editor: José Santamaria
Received: 5 September 2016; Accepted: 13 January 2017; Published: 23 January 2017

Abstract: Human activity recognition (HAR) is an important research area in the fields of human perception and computer vision due to its wide range of applications. These applications include: intelligent video surveillance, ambient assisted living, human computer interaction, human-robot interaction, entertainment, and intelligent driving. Recently, with the emergence and successful deployment of deep learning techniques for image classification, researchers have migrated from traditional handcrafting to deep learning techniques for HAR. However, handcrafted representation-based approaches are still widely used due to some bottlenecks such as computational complexity of deep learning techniques for activity recognition. However, approaches based on handcrafted representation are not able to handle complex scenarios due to their limitations and incapability; therefore, resorting to deep learning-based techniques is a natural option. This review paper presents a comprehensive survey of both handcrafted and learning-based action representations, offering comparison, analysis, and discussions on these approaches. In addition to this, the well-known public datasets available for experimentations and important applications of HAR are also presented to provide further insight into the field. This is the first review paper of its kind which presents all these aspects of HAR in a single review article with comprehensive coverage of each part. Finally, the paper is concluded with important discussions and research directions in the domain of HAR.

Keywords: computer vision; human action recognition; handcrafted representation; learning-based representation; classification; deep learning; Convolutional Neural Networks; review; survey

1. Introduction

In recent years, automatic human activity recognition (HAR) based on computer vision has drawn much attention of researchers around the globe due to its promising results. The major applications of HAR include: Human Computer Interaction (HCI), intelligent video surveillance, ambient assisted living, human-robot interaction, entertainment, and content-based video search. In HCI, the activity recognition systems observe the task carried out by the user and guide him/her to complete it by providing feedback. In video surveillance, the activity recognition system can automatically detect a suspicious activity and report it to the authorities for immediate action. Similarly, in entertainment, these systems can recognize the activities of different players in the game. Depending on the complexity and duration, activities fall into four categories, i.e., gestures, actions, interactions, and group activities [1] as shown in Figure 1.

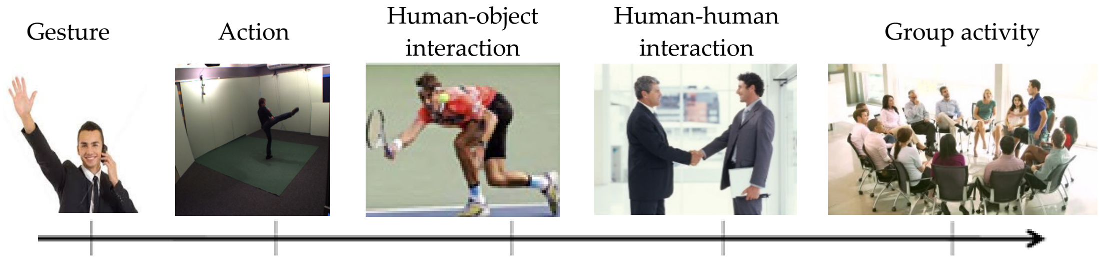


Figure 1. Categorization of different level of activities.

Gesture: A gesture is defined as a basic movement of the human body parts that carry some meaning. 'Head shaking', 'hand waving', and 'facial expression' are some good examples of gestures. Usually, a gesture takes a very short amount of time and its complexity is the lowest among the four mentioned categories.

Action: An action is a type of an activity that is performed by a single person. In fact, it is a combination of multiple gestures (atomic actions). 'Walking' 'running', 'jogging', and 'punching' are some good examples of human actions.

Interaction: It is a type of an activity performed by two actors. One actor must be a human and the other one may be a human or an object. Thus, it could be human-human interaction or human-object interaction. 'Fighting between two persons', 'hand shaking', and 'hugging each other' are examples of human-human interaction, while 'a person using an ATM', 'a person using a computer', 'and a person stealing a bag' are examples of human-object interaction.

Group Activity: This is the most complex type of activity. Certainly, it is a combination of gestures, actions, and interactions. It involves more than two humans and a single or multiple objects. A 'group of people protesting', 'two teams playing a game', and a 'group meeting', are good examples of group activities.

Since the 1980s, researchers have been working on human action recognition from images and videos. One of the important directions that researchers have been following for action recognition is similar to the working of the human vision system. At low level, the human vision system can receive the series of observations regarding the movement and shape of the human body in a short span of time. Then, these observations are passed to the intermediate human perception system for further recognition of the class of these observations, such as walking, jogging, and running. In fact, the human vision and perception system is robust and very accurate in recognition of observed movement and human activities. In order to achieve a similar level of performance by a computer-based recognition system, researchers have carried out a lot of efforts during the past few decades. However, unfortunately, due to many challenges and issues involved in HAR such as environmental complexities, intra-class variations, viewpoint variations, occlusions, and non-rigid shape of the humans and objects, we are still very far from the level of the human vision system. What we have achieved so far may be a fraction of what a mature human vision system can do.

Based on the comprehensive investigation of the literature, vision-based human activity recognition approaches can be divided into two major categories. (1) The traditional handcrafted representation-based approach, which is based on the expert designed feature detectors and descriptors such as Hessian3D, Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), Enhanced Speeded-Up Robust Features (ESURF), and Local Binary Pattern (LBP). This is followed by a generic trainable classifier for action recognition as shown in Figure 2; (2) Learning-based representation approach, which is a recently emerged approach with capability of learning features automatically from the raw data. This eliminates the need of handcrafted feature detectors and descriptors required for action representation as in the traditional approach. Unlike the traditional handcrafted approach, it uses the concept of a trainable feature extractor followed by a trainable classifier, introducing the concept of end-to-end leaning, as shown in Figure 3.

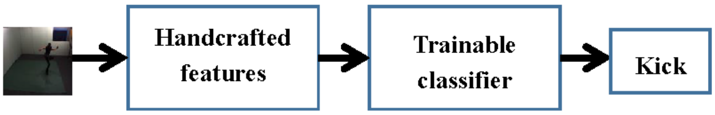


Figure 2. Example of handcrafted representation-based approach.

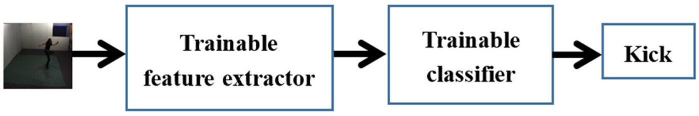


Figure 3. Example of learning-based representation approach.

The handcrafted representation-based approach mainly follows the bottom-up strategy for HAR. Generally, it consists of three major phases (foreground detection, handcrafted feature extraction and representation, and classification) as shown in Figure 4. A good number of survey papers have been published on different phases of handcrafted representation-based HAR processes. Different taxonomies have been used in the survey papers to discuss the HAR approaches. A survey presented in [1], divides the activity recognition approaches into two major categories: single layered approaches and hierarchical approaches. Single-layered approaches recognize the simple activities from the sequence of a video, while hierarchical approaches recognize more complex activities by decomposing them into simple activities (sub-events). These are further sub-categorized such as space-time volumes, and trajectories, based on the feature representation and classification methods used for recognition. A detailed survey on object segmentation techniques is presented in [2], discussing the challenges, resources, libraries and public datasets available for object segmentation. Another study, presented in [3], discussed the three levels of HAR, including core technology, HAR systems, and applications. Activity recognition systems are significantly affected by the challenges such as occlusion, anthropometry, execution rate, background clutter, and camera motion as discussed in [4]. This survey categorized the existing methods based on their abilities for handling these challenges. Based on these challenges, potential research areas were also identified in [4]. In [5], human action recognition methods based on the feature representation and classification were discussed. Similarly, Weinland, D. et al. [6] surveyed the human activity recognition methods by categorizing them into segmentation, feature representation and classification. A review article on semantic-based human action recognition methods is presented in [7]. It presents the state-of-the-art methods for activity recognition which use semantic-based features. In this paper, semantic space, and semantic-based features such as pose, poselet, related objects, attributes, and scene context are also defined and discussed. Different handcrafted features extraction and representation methods have been proposed for human action recognition [8–12].

On the other hand, a learning-based representation approach, specifically deep learning, uses computational models with multiple processing layers based on representation learning with multiple levels of abstraction. This learning encompasses a set of methods that enable the machine to process the data in raw form and automatically transform it into a suitable representation needed for classification. This is what we call trainable feature extractors. This transformation process is handled at different layers, for example, an image consists of an array of pixels, and then the first layer transforms it into edges at particular location and orientation. The second layer represents it as collection of motifs by recognising the particular arrangement of edges in an image. The third layer may combine the motifs into parts and the following layers would turn it into the recognizable objects. These layers are learned from the raw data using a general purpose learning procedure which does not need to be designed manually by the experts [13]. This paper further examines various computer-based fields such as 3D games and animations systems [14,15], physical sciences, health-related issues [16–18], natural sciences and industrial academic systems [19,20].

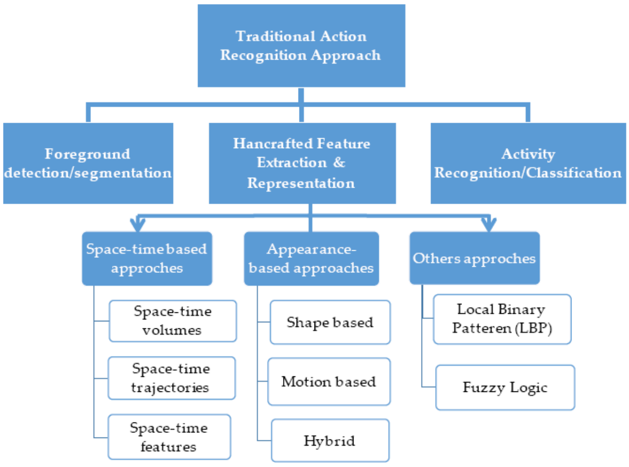

Traditional Action Recognition Approach
Foreground detection/segmentation
Hancrafted Feature Extraction & Representation
Activity Recognition/Classification
Space-time based approches
Appearance-based approaches
Others approches
Space-time volumes
Space-time trajectories
Space-time features
Shape based
Motion based
Hybrid
Local Binary Patteren (LBP)
Fuzzy Logic

Figure 4. Traditional action representation and recognition approach.

One of the important components of vision-based activity recognition system is the camera/sensor used for capturing the activity. The use of appropriate cameras for capturing the activity has great impact on the overall functionality of the recognition system. In fact, these cameras have been instrumental to the progression of research in the field of computer vision [21–25]. According to the nature and dimensionality of images captured by these cameras, they are broadly divided into two categories, i.e., 2D and 3D/depth cameras. The objects in the real world exist in 3D form: when these are captured using 2D cameras then one dimension is already lost, which causes the loss of some important information. To avoid the loss of information, researchers are motivated to use 3D cameras for capturing the activities. For the same reason, 3D-based approaches provide higher accuracy than 2D-based approaches but at higher computational cost. Recently, some efficient 3D cameras have been introduced for capturing images in 3D form. Among these, 3D Time-of-flight (ToF) cameras, and Microsoft Kinect have become very popular for 3D imaging. However, these sensors also have several limitations such as these sensors only capture the frontal surfaces of the human and other objects in the scene. In addition to this, these sensors also have limited range about 6–7 m, and data can be distorted by scattered light from the reflective surfaces [26]. However, there is no universal rule for selecting the appropriate camera; it mainly depends on the nature of the problem and its requirements.

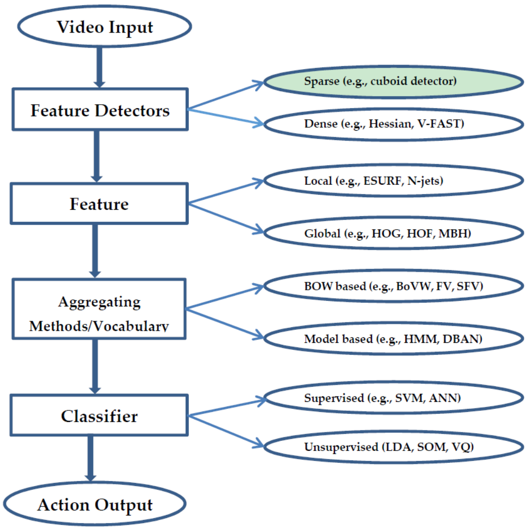
A good number of survey and review papers have been published on HAR and related processes. However, due to the great amount of work about it, already published reviews are always out-of-date. For the same reason, writing a review paper on human activity recognition is hard work and a challenging task. In this paper we provide the discussion, comparison, and analysis of state-of-the-art methods of human activity recognition based on both handcrafted and learning-based action representations along with well-known datasets and important applications. This is the first review article of its kind that covers all these aspects of HAR in a single article with more recent publications. However, this review is more focused towards human gesture, and action recognition techniques, and provides little coverage regarding complex activities such as interactions and group activities. The rest of the paper is organized as follows. Handcrafted representation and recognition-based approaches are covered in Section 2, learning-based representation approaches are discussed in Section 3, Section 4 discusses the well-known public datasets and important application of HAR are presented in Section 5, discussions and conclusion are presented in Section 6.

2. Handcrafted Representation-Based Approach

The traditional approach for action recognition is based on the handcrafted action representation. This approach has been popular among the HAR community and has achieved remarkable results on different public well-known datasets. In this approach, the important features from the sequence of image frames are extracted and the feature descriptor is built up using expert designed feature detectors and descriptors. Then, classification is performed by training a generic classifier such as Support Vector Machine (SVM) [27]. This approach includes space-time, appearance-based, local binary patterns, and fuzzy logic-based techniques as shown in Figure 4.

2.1. Space-Time-Based Approaches

Space-time-based approaches have four major components: space time interest point (STIP) detector, feature descriptor, vocabulary builder, and classifier [28]. The STIP detectors are further categorized into dense and sparse detectors. The dense detectors such as V-FAST, Hessian detector, dense sampling, etc., densely cover all the video content for detection of interest points, while sparse detectors such as cuboid detector , Harris3D [29], and Spatial–Temporal Implicit Shape Model (STISM), etc., use a sparse (local) subset of these contents. Various STIP detectors have been developed by different researchers, such as [30,31]. The feature descriptors are also divided into local and global descriptors. The local descriptors such as cuboid descriptor, Enhanced Speeded-Up Robust Features (ESURF), N-jet are based on the local information such as texture, colour, and posture, while global descriptors use global information such as illumination changes, phase changes, and speed variation of a video. The vocabulary builders or aggregating methods are based on bag-of-words (BOW) or state-space model. Finally, for the classification, a supervised or unsupervised classifier is used, as shown in Figure 5.



Video Input
Feature Detectors
Sparse (e.g., cuboid detector)
Dense (e.g., Hessian, V-FAST)
Feature
Local (e.g., ESURF, N-jets)
Global (e.g., HOG, HOF, MBH)
Aggregating Methods/Vocabulary
BOW based (e.g., BoVW, FV, SFV)
Model based (e.g., HMM, DBAN)
Classifier
Supervised (e.g., SVM, ANN)
Unsupervised (LDA, SOM, VQ)
Action Output

Figure 5. Components of space-time-based approaches. ESURF: Enhanced Speeded-Up Robust Features; HOG: histogram of oriented gradients; HOF: histogram of optical flow; MBH: motion boundary histogram; BOW: bag-of-words; BoVW: Bag-of-Visual-Words; FV: Fisher Vector; SFV: Stacked Fisher Vector; HMM: Hidden Markov Model; DBAN: Dynamic Bayesian Action Network; SVM: support vector machine; ANN: Artificial Neural Network; LDA: Latent Dirichlet Allocation; SOM: Self Organizing Map; VQ: vector quantization.

2.1.1. Space-Time Volumes (STVs)

The features in the space-time domain are represented as 3D spatio-temporal cuboids, called space-time volumes (STVs). The core of STV-based methods is similarity measure between two volumes for action recognition. In [32], an action recognition system was proposed using template matching, instead of using space-time volumes, they used templates composed of 2D binary motion-energy-image (MEI) and motion-history-image (MHI) for action representation followed by a simple template matching technique for action recognition. This work was extended in [33] where MHI and two appearance-based features namely foreground image and histogram of oriented gradients (HOG) were combined for action representation, followed by simulated annealing multiple instance learning support vector machine (SMILE-SVM) for action classification. The method proposed in [34] also extended the [32] from 2D to 3D space for view-independent human action recognition using a volume motion template. The experimental results using these techniques are presented in Table 1.

Table 1. Comparison of Space-Time-based approaches for activity recognition on different datasets.

Method	Feature Type	Performance (%)
	KTH [35]	
Sadanand and Corso 2012 [36]	Space-time volumes	98.2
Wu et al. 2011 [37]	Space-time volumes	94.5
Ikizler and Duygulu 2009 [38]	Space-time volumes	89.4
Peng et al. 2013 [39]	Features	95.6%
Liu et al. 2011 [40]	Features (Attributes)	91.59
Chen et al. 2015 [41]	Features (mid-level)	97.41
Wang et al. 2011 [42]	Dense trajectory	95%
	UCF (University of Central Florida) Sports [43,44]	
Sadanand and Corso 2012 [36]	Space-time volumes	95.0
Wu et al. 2011 [37]	Space-time volumes	91.30
Ma et al. 2015 [45]	Space-time volumes	89.4
Chen et al. 2015 [41]	Features (mid-level)	92.67
Wang et al. 2013 [46]	Features (Pose-based)	90
Sadanand and Corso 2012 [36]	STVs (space-time volumes)	95.0
	HDMB (Human Motion database)-51 [47]	
Wang and Schmid 2013 [48]	Dense trajectory	57.2
Jiang et al. 2012 [49]	Trajectory	40.7
Wang et al. 2011 [42]	Dense trajectory	46.6
Kliper et al. 2012 [50]	Space-time volumes, bag-of-visual-words	29.2
Sadanand and Corso 2012 [36]	Space-time volumes	26.9
Kuehne et al. 2011 [47]	Features	23.0
Wang et al. 2013 [51]	Features (mid-level)	33.7
Peng et al. 2014 [52]	Fisher vector and Stacked Fisher Vector	66.79
Jain et al. 2013 [53]	Features	52.1
Fernando et al. 2015 [54]	Features (Video Darwin)	63.7
Hoai and Zisserman 2014 [55]	Features	65.9
	Hollywood2 [56]	
Wang and Schmid 2013 [48]	Dense trajectory	64.3
Jain et al. 2013 [53]	Trajectory	62.5
Jiang et al. 2012 [49]	Trajectory	59.5
Vig et al. 2012 [57]	Trajectory	59.4
Mathe and Sminchisescu 2012 [58]	Space-time volumes	61.0
Kihl et al. 2016 [59]	Features	58.6
Lan et al. 2015 [60]	Features (mid-level)	66.3
Fernando et al. 2015 [54]	Features (Video Darwin)	73.7
Hoai and Zisserman [55]	Features	73.6
	Microsoft Research Action3D [61]	
Wang et al. 2013 [46]	Features (pose-based)	90.22
Amor et al. 2016 [62]	Trajectory	89
Zanfir et al. 2013 [63]	3D Pose	91.7
	YouTube action dataset [64]	
Wang et al. 2011 [42]	Dense trajectory	84.1
Peng et al. 2014 [52]	Features (FV + SFV)	93.38

2.1.2. Space-Time Trajectory

Trajectory-based approaches interpret an activity as a set of space-time trajectories. In these approaches, a person is represented by 2-dimensional (*XY*) or 3-dimensional (*XYZ*) points corresponding to his/her joints position of the body. As person performs an action, there are certain changes in his/her joint positions according to the nature of the action. These changes are recoded as space-time trajectories, which construct a 3D *XYZ* or 4D *XYZT* representation of an action. The space-time trajectories work by tracking the joint position of the body for distinguishing different types of actions. Following this idea many approaches have been proposed for action recognition based on the trajectories [46,65,66].

Inspired by the dense sampling in image classification, the concept of dense trajectories for action recognition from videos was introduced in [42]. The authors sampled the dense points from each image frame and tracked them using displacement information from a dense optical flow field. These types of trajectories cover the motion information and are robust to irregular motion changes. This method achieved state-of-the-art-results on challenging datasets. In [48], an extension to [42] was proposed for the improvement of performance regarding camera motion. For estimation of camera motion authors used Speeded-Up Robust Features (SURF) descriptor and dense optical flow. This significantly improved the performance of motion-based descriptors such as histogram of optical flow (HOF), and motion boundary histogram (MBH). However, when incorporating high density with trajectories within the video, it increases the computational cost. Many attempts have been made to reduce the computational cost of the dense trajectory-based methods. For this purpose, saliency-map to extract the salient regions within the image frame was used in [57]. Based on the saliency-map a significant number of dense trajectories can be discarded without compromising the performance of the trajectory-based methods.

Recently, a human action recognition method from depth movies captured by the Kinect sensor was proposed in [62]. This method represents dynamic skeleton shapes of the human body as trajectories on Kendall's shape manifold. This method is invariant to execution rate of the activity and uses transported-square root vector fields (TSRVFs) of trajectories and standard Euclidean norm to achieve the computational efficiency. Another method for recognition of actions of construction workers using dense trajectories was proposed in [67]. In this method, different descriptors such as HOG, HOF, and motion boundary histogram (MBH) were used for the trajectories. Among these descriptors, authors reported the highest accuracy with codebook of size 500 using MBH descriptor. Human action recognition in unconstrained videos is a challenging problem and few methods have been proposed at this end. For this purpose, a human action recognition method was proposed using explicit motion modelling [68]. This method used visual code words generated from the dense trajectories for action representation without using the foreground-background separation method.

2.1.3. Space-Time Features

The space-time features-based approaches extract features from space-time volumes or space-time trajectories for human action recognition. Generally, these features are local in nature and contain discriminative characteristics of an action. According to the nature of space-time volumes and trajectories, these features can be divided into two categories: sparse and dense. The features detectors that are based on interest point detectors such as Harris3D [30], and Dollar [69] are considered as sparse, while feature detectors based on optical flow are considered as dense. These interest point detectors provide the base for most of the recently proposed algorithms. In [70], the interest points were detected using Harris3D [30], based on these points they build the feature descriptor and used PCA (principal component analysis)-SVM for classification. In [59], authors proposed a novel local polynomial space-time descriptor based on optical flow for action representation.

The most popular action representation methods in this category are based on the Bag-of-Visual-Words (BoVW) model [71,72] or its variants [73,74]. The BoVW model consists of four steps, feature extraction, codebook generation, encoding and pooling, and normalization. We extract the local features from the video; learn visual dictionary in training set by Gaussian Mixture Model (GMM) or K-mean clustering, encode and pool features, and finally represent the video as normalized pooled vectors followed by a generic classifier for action recognition. The high performance of the BoVW model is due to an effective low level feature such as dense trajectory features [48,75], encoding methods such as Fisher Vector [74], and space-time co-occurrence descriptors [39]. The improved dense trajectory (iDT) [48] provides the best performance among the space-time features on serval public datasets.

The coding methods have played an important role in boosting the performance of these approaches. Recently, a new encoding method named Stacked Fisher Vector (SFV) [52] was developed as an extension of traditional single layer Fisher Vector (FV) [74]. Unlike traditional FV, which encodes all local descriptors at once, SFV first performs encoding in dense sub-volumes, then compresses these sub-volumes into FVs, and finally applies another FV encoding based on the compressed sub-volumes. For the detail comparison of single layer FV and stacked FV readers are encouraged to refer [52].

2.1.4. Discussion

Space-Time-based approaches have been evaluated by many researchers on different well-known datasets—including simple and complex activities as recoded in Table 1. Some merits of these approaches are as follows: (1) STVs-based approaches are suitable for recognition of gestures and simple actions. However, these approaches have also produced comparable results on complex datasets such as Human Motion database (HMDB-51), Hollywood2, and University of Central Florida (UCF-101); (2) The space-time trajectory-based approaches are especially useful for recognition of complex activities. With the introduction of dense trajectories, these approaches have become popular due to their high accuracy for challenging datasets. In recent years, trajectory-based approaches are getting lot of attention due to their reliability under noise and illumination changes; (3) Space-time feature-based approaches have achieved state-of-the art results on many challenging datasets. It has been observed that descriptors such as HOG3D, HOG/HOF, and MBH are more suitable for handling intra-class variations and motion challenges in complex datasets as compared to local descriptors such as N-jet.

However, these approaches have some limitations as follows: (1) STVs-based approaches are not effective in recognising multiple persons in a scene; these methods use a sliding window for this purpose which is not very effective and efficient; (2) Trajectory-based approaches are good at analysing the movement of a person in view invariant manner but to correctly localize the 3D XYZ joint position of a person is still a challenging task; (3) Space-time features are more suitable for simple datasets; for effective results on complex datasets, combination of different features is required which raises the computational complexity. These limitations can cause hindrance for real-time applications.

2.2. Appearance-Based Approaches

In this section we discuss the 2D (XY) and 3D (XYZ) depth image-based approaches which use effective shape, motion, or combination of shape and motion features for action recognition. The 2D shape-based approaches [76,77] use shape and contour-based features for action representation and motion-based approaches [78,79] use optical flow or its variants for action representations. Some approaches use both shape and motion feature for action representation and recognition [80]. In 3D-based approaches, a model of a human body is constructed for action representation; this model can be based on cylinders, ellipsoids, visual hulls generated from silhouettes or surface mesh. Some examples of these methods are 3D optical flow [81], shape histogram [82], motion history volume [83], and 3D body skeleton [84].

2.2.1. Shape-Based Approaches

The shape-based approaches capture the local shape features from the human image/silhouette [85]. These methods first obtained the foreground silhouette from an image frame using foreground segmentation techniques. Then, they extract the features from the silhouette itself (positive space) or from the surrounding regions of the silhouette (negative space) between canvas and the human body [86]. Some of the important features that can be extracted from the silhouette are contour points, region-based features, and geometric features. The region-based human action recognition method was proposed in [87]. This method divides the human silhouette into a fixed number of grids and cells for action representation and used a hybrid classifier Support Vector Machine and Nearest Neighbour (SVM-NN) for action recognition. For practical applications, the human action recognition method should be computationally lean. In this direction, an action recognition method was proposed using Symbolic Aggregate approximation (SAX) shapes [88]. In this method, a silhouette was transformed into time-series and these time-series were converted into a SAX vector for action representation followed by a random forest algorithm for action recognition.

In [89], a pose-based view invariant human action recognition method was proposed based on the contour points with sequence of the multi-view key poses for action representation. An extension of this method was proposed in [90]. This method uses the contour points of the human silhouette and radial scheme for action representation and support vector machine as a classifier. In [86], and [91] a region-based descriptor for human action representation was developed by extracting features from the surrounding regions (negative space) of the human silhouette. Another method used pose information for action recognition [92]. In this method, first, the scale invariant features were extracted from the silhouette, and then these features were clustered to build the key poses. Finally, the classification was performed using a weighted voting scheme.

2.2.2. Motion-Based Approaches

Motion-based action recognition approaches use the motion features for action representation followed by a generic classifier for action recognition. A novel motion descriptor was proposed in [93] for multi-view action representation. This motion descriptor is based on motion direction and histogram of motion intensity followed by the support vector machine for classification. Another method based on 2D motion templates using motion history images and histogram of oriented gradients was proposed in [94]. In [50], action recognition method was proposed based on the key elements of motion encoding and local changes in motion direction encoded with the bag-of-words technique.

2.2.3. Hybrid Approaches

These approaches combine shape-based and motion-based features for action representation. An optical flow and silhouette-based shape features were used for view invariant action recognition in [95] followed by principal component analysis (PCA) for reducing the dimensionality of the data. Some other methods based on shape and motion information were proposed for action recognition in [96,97]. The coarse silhouette features, radial grid-based features and motion features were used for multi-view action recognition in [97]. Meanwhile, [80] used shape-motion prototype trees for human action recognition. The authors represented action as a sequence of porotypes in shape-motion space and used distance measure for sequence matching. This method was tested on five public datasets and achieved state-of-the-art results. In [98], the authors proposed a method based on action key poses as a variant of Motion Energy Images (MEI), and Motion History Images (MHI) for action representation followed by simple nearest-neighbour classifier for action recognition.

2.3. Other Approaches

In this section we discuss the two important approaches that do not fit under the headings of above mentioned categories. These approaches include Local Binary Pattern (LBP), and fuzzy logic-based methods.

2.3.1. Local Binary Pattern (LBP)-Based Approaches

Local binary patterns (LBP) [99] is a type of visual descriptor for texture classification. Since its inception, several modified versions of this descriptor such as [100–102] have been proposed for different classification-related tasks in computer vision. A human action recognition method was proposed in [103] based on LBP combined with appearance invariance and patch matching method. This method was tested on different public datasets and proved to be efficient for action recognition. Another method for activity recognition was proposed using LBP-TOP descriptor [104]. In this method, the action volume was partitioned into sub-volumes and feature histogram was generated by concatenating the histograms of sub-volumes. Using this representation, they encoded the motion at three different levels: pixel-level (single bin in the histogram), region-level (sub-volume histogram), and global-level (concatenation of sub-volume-histograms). The LBP-based methods have also been employed for multi-view human action recognition. In [105], a multi-view human action recognition method was proposed based on contour-based pose features and uniform rotation-invariant LBP followed by SVM for classification. Recently, another motion descriptor named Motion Binary Pattern (MBP) was introduced for multi-view action recognition [106]. This descriptor is combination of Volume Local Binary Pattern (VLBP) and optical flow. This method was evaluated on multi-view INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset and achieved 80.55% recognition results.

2.3.2. Fuzzy Logic-Based Approaches

Traditional vision-based human action recognition approaches employ the spatial or temporal features followed by a generic classifier for action representation and classification. However, it is difficult to scale up these approaches for handling uncertainty and complexity involved in real world applications. For handling these difficulties, fuzzy-based approaches are considered as a better choice. In [107], a fuzzy-based framework was proposed for human action recognition based on the fuzzy log-polar histograms and temporal self-similarities for action representation followed by SVM for action classification. The evaluation of proposed method on two public datasets confirmed the high accuracy and its suitability for real world applications. Another method based on fuzzy logic was proposed in [108]. This method utilized the silhouette slices and movement speed features as input to the fuzzy system, and employed the fuzzy c-means clustering technique to acquire the membership function for the proposed system. The results confirmed the better accuracy of the proposed fuzzy system as compared to non-fuzzy systems for the same public dataset.

Most of the human action recognition methods are view dependent and can recognize the action from a fixed view. However, a real-time human action recognition method must be able to recognize the action from any viewpoint. To achieve this objective, many state-of-the-art methods use multi-camera setup in their processing. However, this is not a practical solution because calibration of multiple cameras in real world scenarios is quite difficult. The use of a single camera should be the ultimate solution for view invariant action recognition. Along these lines, a fuzzy logic-based method was proposed in [109] for view invariant action recognition using a single camera. This method extracted human contour from fuzzy qualitative Poisson human model for view estimation followed by clustering algorithms for view classification as shown in Figure 6. The results indicate that the proposed method is quite efficient for view independent action recognition. Some methods based on neuro-fuzzy systems (NFS) have also been proposed for human gesture and action recognition [110,111]. In addition to this, evolving systems [112,113] are also very successful in behaviour recognition.

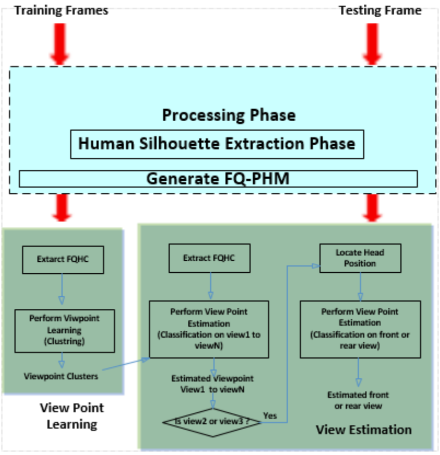

Training Frames
Testing Frame
Processing Phase
Human Silhouette Extraction Phase
Generate FQ-PHM
Extarct FQHC
Perform Viwpoint Learning (Clustring)
Viewpoint Clusters
View Point Learning
Extract FQHC
Perform View Point Estimation (Classification on view1 to viewN)
Estimated Viewpoint View1 to viewN
Is view2 or view3 ?
Yes
Locate Head Position
Perform View Point Estimation (Classification on front or rear view)
Estimated front or rear view
View Estimation

Figure 6. Example of Fuzzy view estimation framework.

2.3.3. Discussion

In this section we compare the appearance, LBP and fuzzy logic-based methods. These approaches are simple and have produced state-of-the-art results on Weizmann, KTH, and multi-view IXMAS datasets as recorded in Table 2. There are two major approaches for multi-view human action recognition base on shape and motion features: 3D approach and 2D approach [26]. As indicated in Table 2, 3D approaches provide higher accuracy than 2D approaches but at higher computational cost which makes these approaches less applicable for real time applications. In addition to this, it is difficult to reconstruct a good quality 3D model because it depends on the quality of extracted features or silhouettes of different views. Hence, the model is exposed to deficiencies which might have occurred due to segmentation errors in each view point. Moreover, a good 3D model of different views can only be constructed when the views overlap. Therefore, a sufficient number of viewpoints have to be available to reconstruct a 3D model.

Table 2. Comparison of appearance, LBP (Local Binary Pattern), and fuzzy logic-based approaches.

Method	Feature Type	Performance (%)
Weizmann [114]		
Rahman et al. 2012 [86]	Shape Features	100
Vishwakarma and Kapoor 2015 [87]	Shape Features	100
Rahman et al. 2014 [91]	Shape-motion	95.56
Chaaraoui et al. 2013 [89]	Shape Features	92.8
Vishwakarma et al. 2016 [96]	Shape Features	100
Jiang et al. 2012 [80]	Shape-motion	100
Eweiwi et al. 2011 [98]	Shape-motion	100
Yeffet and Wolf 2009 [103]	LBP	100
Kellokumpu et al. 2008 [104]	LBP (LBP-TOP)	98.7
Kellokumpu et al. 2011 [115]	LBP	100
Sadek et al. 2011 [107]	Fuzzy features	97.8
Yao et al. 2015 [108]	Fuzzy features	94.03
KTH [35]		
Rahman et al. 2012 [86]	Shape Features	94.67
Vishwakarma and Kapoor 2015 [87]	Shape Features	96.4
Rahman et al. 2014 [91]	Shape-motion	94.49
Vishwakarma et al. 2016 [96]	Shape Features	95.5
Sadek et al. 2012 [116]	Shape Features	93.30
Jiang et al. 2012 [80]	Shape-motion	95.77
Yeffet and Wolf 2009 [103]	LBP	90.1
Mattivi and Shao 2009 [117]	LBP (LBP-TOP)	91.25
Kellokumpu et al. 2011 [115]	LBP	93.8
Sadek et al. 2011 [107]	Fuzzy Features	93.6
IXMAS (INRIA Xmas Motion Acquisition Sequences) [118]		
Junejo et al. 2014 [88]	Shape Features	89.0
Sargano et al. 2016 [119]	Shape features	89.75
Lin et al. 2009 [80]	Shape-motion	88.89
Chaaraoui et al. 2013 [89]	Shape Features	85.9
Chun and Lee 2016 [93]	Motion Features	83.03
Vishwakarma et al. 2016 [96]	Shape Features	85.80
Holte et al. 2012 [120]	Motion Feature (3D)	100
Weinland et al. 2006 [83]	Motion Features (3D)	93.33
Turaga et al. 2008 [121]	Shape-motion (3D)	98.78
Pehlivan and Duygulu 2011 [122]	Shape Features (3D)	90.91
Baumann et al. 2016 [106]	LBP	80.55

3. Learning-Based Action Representation Approach

The performance of the human action recognition methods mainly depends on the appropriate and efficient representation of data. Unlike handcrafted representation-based approaches where the action is represented by handcrafted feature detectors and descriptors; learning-based representation approaches have capability to learn the feature automatically from the raw data, thus introducing the concept of end-to-end learning which means transformation from pixel level to action classification. Some of these approaches are based on evolutionary approach (genetic programming) and dictionary learning while others employ deep learning-based models for action representation. We have divided these approaches into two categories: non-deep learning-based approaches and deep learning-based approaches as shown in Figure 7.

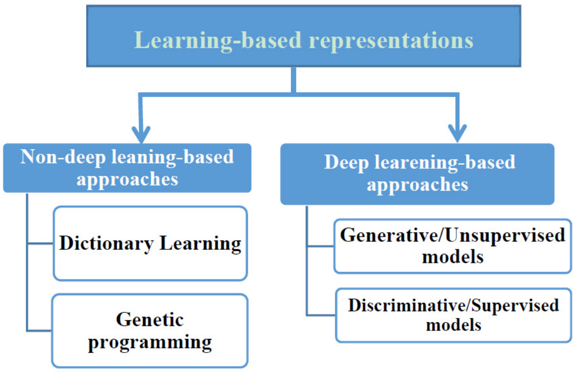

Learning-based representations
Non-deep leaning-based approaches
Deep learening-based approaches
Dictionary Learning
Genetic programming
Generative/Unsupervised models
Discriminative/Supervised models

Figure 7. Learning-based action representation approaches.

3.1. Non-Deep Learning-Based Approaches

These approaches are based on genetic programming and dictionary learning as discusses in the following section.

3.1.1. Dictionary Learning-Based Approaches

Dictionary learning is a type of representation learning which is generally based on the sparse representation of input data. The sparse representation is suitable for the categorization tasks in images and videos. Dictionary learning-based approaches have been employed in wide range of computer vision applications such as image classification and action recognition [123]. The concept of dictionary learning is similar to BoVW model because both are based on the representative vectors learned from the large number of samples. These representative vectors are called code words, forming a codebook in BoVW model, and dictionary atoms in the context of dictionary learning. One way to get the sparse representation of input data is to learn over-complete basis (dictionary). In [124], three over-complete dictionary learning frameworks were investigated for human action recognition. An over-complete dictionary was constructed from a set of spatio-temporal descriptors, where each descriptor was represented by a linear combination of small number of dictionary elements for compact representation. A supervised dictionary learning-based method was proposed for human action recognition in [125] based on the hierarchical descriptor. Cross-view action recognition problem was addressed by using transferable dictionary pair in [126]. In this approach authors learned the view specific dictionaries where each dictionary corresponds to one camera view. Moreover, authors extended this work and a common dictionary which shares information from different views [127]. The proposed approach outperforms state-of-the-art methods on similar datasets. A weakly supervised cross-domain dictionary learning-based method was proposed for visual recognition in [128]. This method learns discriminative, domain-adaptive, and reconstructive dictionary pair and corresponding classifier parameters without any prior information.

Dictionary leaning-based methods also use unsupervised learning, for example, Zhu, F. et al. [129] proposed an unsupervised approach for cross-view human action recognition. This method does not require target view label information or correspondence annotations for action recognition. The set of low-level trajectory features are coded using locality-constrained linear coding (LLC) [130] to form the coding descriptors, then peak values are pooled to form a histogram that captures the local structure of each action.

3.1.2. Genetic Programming

Genetic programming is a powerful evolutionary technique inspired by the process of natural evolution. It can be used to solve the problems without having the prior knowledge of solutions. In human activity, recognition genetic programming can be employed to identify the sequence of unknown primitive operations that can maximize the performance of recognition task. Recently, a genetic programming-based approach was introduced for action recognition in [131]. In this method, instead of using handcrafted features, authors automatically learned the spatio-temporal motion features for action recognition. This motion feature descriptor evolved on population of 3D operators such as 3D-Gabor filter and wavelet. In this way, effective features were learnt for action recognition. This method was evaluated on three challenging datasets and outperformed the handcrafted as well as other learning-based representations.

3.2. Deep Learning-Based Approaches

Recent studies show that there are no universally best hand-crafted feature descriptors for all datasets, therefore learning features directly from the raw data may be more advantageous. Deep learning is an important area of machine learning which is aimed at learning multiple levels of representation and abstraction that can make sense of data such as speech, images, and text. These approaches have the ability to process the images/videos in their raw forms and automate the process of feature extraction, representation, and classification. These approaches use trainable feature extractors and computational models with multiple processing layers for action representation and recognition. Based on the research study on deep learning presented in [132], we have classified the deep learning models into three categories: (1) generative/unsupervised models (e.g., Deep Belief Networks (DBNs), Deep Boltzmann machines (DBMs), Restricted Boltzmann Machines (RBMs), and regularized auto-encoders); (2) Discriminative/Supervised models (e.g., Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs); (3) Hybrid models, these models use the characteristics of both models, for example a goal of discrimination may be assisted with the outcome of the generative model. However, we are not going to discuss hybrid models separately rather discuss these either in the supervised or unsupervised category.

3.2.1. Generative/Unsupervised Models

Generative/unsupervised deep learning models do not require the target class labels during the learning process. These models are specifically useful when labelled data are relatively scarce or unavailable. Deep learning models have been investigated since the 1960s [133] but researchers have paid little attention towards these models. This was mainly due to the success of shallow models such as SVMs [134], and unavailability of huge amount of data, required for training the deep models.

A remarkable surge in the history of deep models was triggered out by the work of [135] where the highly efficient DBN and training algorithm was introduced followed by the feature reduction technique [136]. The DBN was trained layer by layer using RBMs [137], the parameters learned during this unsupervised pre-training phase were fine-tuned in a supervised manner using backpropagation. Since the introduction of this efficient model there has been a lot of interest in applying deep learning models to different applications such as speech recognition, image classification, object recognition, and human action recognition.

A method using unsupervised feature learning from video data was proposed in [138] for action recognition. The authors used independent subspace analysis algorithm for learning spatio-temporal features and combined with deep learning techniques such as convolutional and staking for action representation and recognition. Deep Belief Networks (DBNs) trained with RBMs were used for human action recognition in [139]. The proposed approach outperforms the handcrafted learning-based approaches on two standard datasets. Learning continuously from the streaming video without any labels is an important but challenging task. This issue was addressed in [140] by using unsupervised deep learning model. Most of action datasets have been recoded under controlled environment; action recognition from unconstrained videos is a challenging task. A method for human action recognition from unconstrained video sequences was proposed in [141] using DBNs.

Unsupervised learning played a pivotal role in reviving the interests of the researchers in deep learning. However, it has been overshadowed by the purely supervised learning since the major breakthrough in deep learning used CNNs for object recognition [142]. However, an important study by the pioneers of latest deep learning models suggest that unsupervised learning is going to be far more important than its supervised counterpart in the long run [13] since we discover the world by observing it rather being told the name of every object. The human and animal learning is mostly unsupervised.

3.2.2. Discriminative/Supervised Models

According to the literature survey of human action recognition, the most frequently used model under the supervised category is Convolutional Neural Networks (CNN or Convnet). The CNN [143] is a type of deep learning model which has shown excellent preformation at tasks such as pattern recognition, hand-written digit classification, image classification and human action recognition [142,144]. This is a hierarchical learning model with multiple hidden layers to transform the input volume into output volume. Its architecture consists of three main types of layers: convolutional layer Convolution and Rectifier Linear Unit (CONV + ReLU), pooling layer, and fully-connected layer as shown in Figure 8. Understanding the operation of different layers of CNN require mapping back these activities into pixel space, this is done with the help of Deconvolutional Networks (Deconvnets) [145]. The Deconvnets use the same process as CNN but in reverse order for mapping from feature space to pixel space.

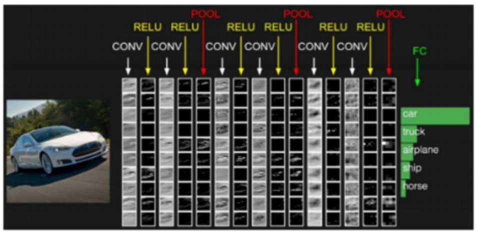


Figure 8. Convolutional Neural Networks layers (source [146]).

Initially, the deep CNN [143] was used for representation and recognition of objects from still images [142]. This was extended to action recognition from videos in [147] using stacked video frames as input to the network but the results were worse than even the handcrafted shallow representations [48,72]. This issue was investigated in [148] and they came up with the idea of two-stream (spatial and temporal) CNN for action recognition. An example of two-stream convolutional neural network is shown in Figure 9. These both streams were implemented as Convnet, the spatial stream recognizes the action from still video frames and the temporal stream performs action recognition from the motion in the form of dense optical flow. Afterwards, these two streams were combined using late fusion for action recognition. This method achieved superior results to one of the best shallow handcrafted-based representation methods [48]. However, the two-stream architecture may not be suitable for real-time applications due to its computational complexity.

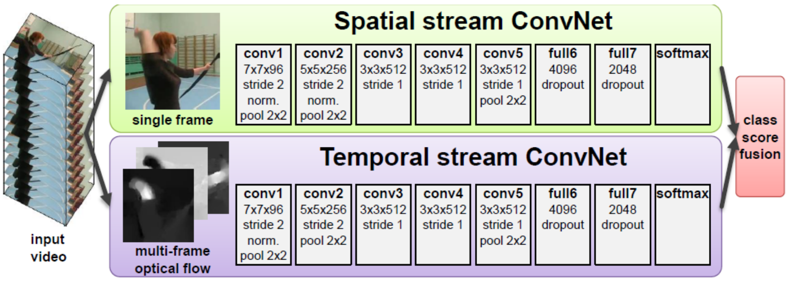


Figure 9. Two-stream Convolutional Neural Network (CNN) architecture (Source [148]).

Most of the deep CNN models for action recognition are limited to handling inputs in 2D form. However, some applications do have data in 3D form that requires a 3D CNN model. This problem was addressed in [149] by introducing the 3D convolutional neural networks model for the airport surveillance. This model uses features from both spatial and temporal dimensions by performing 3D convolutions at the convolutional layer. This method achieved state-of-the-art results in airport video surveillance datasets.

The supervised learning CNN model 2D or 3D can also be accompanied by some unsupervised endeavours. One of the unsupervised endeavours is slow feature analysis (SFA) [150], which extracts slowly varying features from the input signal in an unsupervised manner. Beside other recognition problems, it has proved to be effective for human action recognition as well [151]. In [152], two-layered SFA learning was combined with 3D CNN for automated action representation and recognition. This method achieved state-of-the-art results on three public datasets including KTH, UCF sports, and Hollywood2. Other types of supervised models include Recurrent Neural Networks (RNNs). A method using RNN was proposed in [153] for skeleton-based action recognition. The human skeleton was divided into five parts and then separately fed into five subnets. The results of these subnets were fused into the higher layers and final representation was fed into the single layer. For further detail regrading this model reader may refer to [153].

The deep learning-based models for human action recognition require huge amount of video data for training. However, collecting and annotating huge amount of video data is immensely laborious and requires huge computational resources. A remarkable success has been achieved in the domains of image classification, object recognition, speech recognition, and human action recognition using the standard 2D and 3D CNN models. However, still there exist some issues such as high computational complexity of training CNN kernels, and huge data requirements for training. To curtail these issues researchers have been working to come up with variations/adaptations of these models. In this direction, factorized spatio-temporal convolutional networks (FSTCN) was proposed in [154] for human action recognition. This network factorizes the standard 3D CNN model as a 2D spatial kernels at lower layers (spatial convolutional layers) based on sequential learning process and 1D temporal kernels in the upper layers (temporal convolutional layers). This reduced the number of parameters to be learned by the network and thus reduced the computational complexity of the training CNN kernels. The detailed architecture is presented in [154].

Another approach using spatio-temporal features with a 3D convolutional network was proposed in [155] for human action recognition. The evaluation of this method on four public datasets confirmed three important findings: (1) 3D CNN is more suitable for spatio-temporal features than 2D CNN; (2) The CNN architecture with small 3 × 3 × 3 kernels is the best choice for spatio-temporal features; (3) The proposed method with linear classifier outperforms the state-of-the-art methods.

Some studies have reported that incorporating handcrafted features into the CNN model can improve the performance of action recognition. Along this direction, combining information from multiple sources with CNN was proposed in [156]. The authors used handcrafted features to perform spatially varying soft-gating and used fusion method for combining multiple CNNs trained on different sources. Recently, another variation of CNN was proposed in [157], called stratified pooling-based CNN (SP-CNN). Since each video has a different number of frame-level features, to combine and get a video-level feature is a challenging task. The SP-CNN method addressed this issue by proposing variation in the CNN model as follows: (a) adjustment of pre-trained CNN on target dataset; (b) extraction of features at frame-level; (c) using principal component analysis (PCA) for dimensionality reduction; (d) stratified pooling frame-level features into video-level features; (e) SVM for multiclass classification. This architecture is shown in Figure 10.

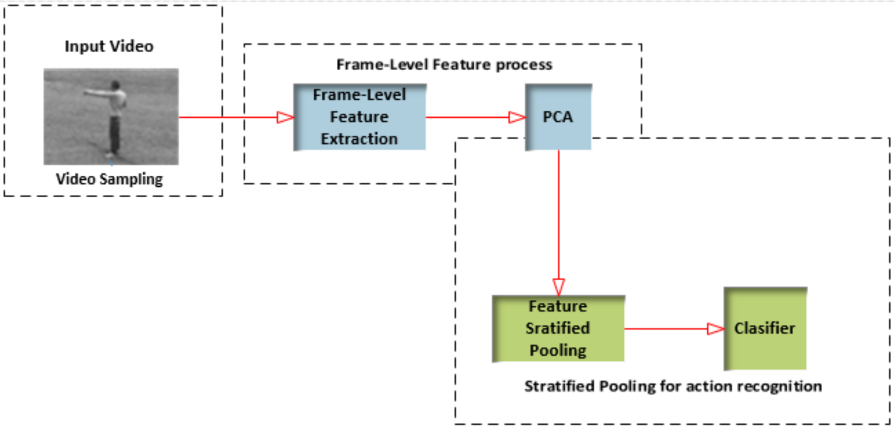

Input Video
Video Sampling
Frame-Level Feature process
Frame-Level Feature Extraction
PCA
Feature Sratified Pooling
Clasifier
Stratified Pooling for action recognition

Figure 10. An example of stratified pooling with CNN.

Sematic-based features such as pose, poselet are important cues for describing the category of an action being performed. In this direction, some methods based on fuzzy CNN were proposed in [158,159] using local posed-based features. These descriptors are based on the motion and appearance information acquired from tracking human body parts. These methods were evaluated on Human Motion Database (HMDB) produced superior results than other state-of-the-art methods. It has been observed that the context/scene where the action is carried out also provides important cues regarding the category of an action. In [160], the contextual information was exploited for human action recognition. They adapted the Region-based Convolutional Neural Network (RCNN) [161] to use more than one region for classification. It considered the actor as a primary region and contextual cues as a secondary region.

One of the major challenges in human action recognition is view variance. The same action viewed from different angles looks quite different. This issue was addressed in [162] using CNN. This method generates the training data by fitting synthetic 3D human model to real motion and renders human poses from different viewpoints. Convolutional Neural Networks model has shown better performance than handcrafted representation-based methods for multi-view human action recognition. Table 3 shows the comparison of non-deep learning and deep learning-based methods on different public datasets.

Table 3. Comparison of Learning-based action representation approaches.

Method	Feature Type	Performance (%)
KTH [35]		
Wang et al. 2012 [125]	Dictionary Learning	94.17
Liu et al. 2016 [131]	Genetic Programming	95.0
Le et al. 2011 [138]	Subspace analysis	93.9
Ballan et al. 2012 [141]	Codebook	92.66
Hasan and Chowdhury 2014 [140]	DBNs (Deep Belief Networks)	96.6
Ji et al 2013 [149]	3D CNN (Convolutional Neural Networks)	90.2
Zhang and Tao 2012 [151]	Slow Feature Analysis (SFA)	93.50
Sun et al. 2014 [152]	Deeply-Learned Slow Feature Analysis (D-SFA)	93.1
Alfaro et al. 2016 [163]	Sparse coding	97.5%
HDMB-51 [47]		
Liu et al. 2016 [131]	Genetic Programming	48.4
Simonyan and Zisserman 2014 [148]	CNN	59.4
Luo et al. 2015 [164]	Actionness	56.38
Wang et al. 2015 [165]	convolutional descriptor	65.9
Lan et al. 2015 [166]	Multi-skip Feature Stacking	65.1
Sun et al. 2015 [154]	Spatio-Temporal CNN	59.1
Park et al. 2016 [156]	Deep CNN	54.9
Yu et al. 2016 [157]	SP(stratified pooling)-CNN	74.7
Bilen et al. 2016 [167]	Multiple Dynamic Images (MDI), trajectory	65.2
Mahasseni and Todorovic 2016 [168]	Lon Short term Memory- Convolutional Neural Network (LSTM-CNN)	55.3
Fernando et al. 2016 [169]	Rank pooling + CNN	65.8
Zhu et al. 2016 [170]	Key volume mining	63.3
Hollywood2 [56]		
Liu et al. 2016 [131]	Genetic Programming	46.8
Le et al. 2011 [138]	Subspace analysis	53.3
Ballan et al. 2012 [141]	Codebook	45.0
Sun et al. 2014 [152]	DL-SFA	48.1
Fernando et al. 2016 [169]	Rank pooling + CNN	75.2
MSR Action3D [61]		
Du et al. 2015 [153]	RNN (Recurrent Neural Network)	94.49
Wang et al. 2016 [171]	3D Key-Pose-Motifs	99.36
Veeriah et al. 2015 [172]	Differential RNN	92.03
University of Central Florida (UCF-101) [173]		
Simonyan and Zisserman 2014 [148]	Two-stream CNN	88.0
Ng et al. 2015 [174]	CNN	88.6
Wang et al. 2015 [165]	convolutional descriptor	91.5
Lan et al. 2015 [166]	Multi-skip Feature Stacking	89.1
Sun et al. 2015 [154]	Spatio-Temporal CNN	88.1
Tran et al. 2015 [155]	3D CNN	90.4
Park et al. 2016 [156]	Deep CNN	89.1
Yu et al. 2016 [157]	SP-CNN	91.6
Bilen et al. 2016 [167]	MDI and trajectory	89.1
Mahasseni and Todorovic 2016 [168]	LSTM-CNN	86.9
Zhu et al. 2016 [170]	Key volume mining	93.1
UCF Sports [43,44]		
Sun et al. 2014 [152]	DL-SFA	86.6
Weinzaepfel et al. 2015 [175]	Spatio-temporal	91.9%
ActivityNet Dataset [176]		
Heilbron et al. 2015 [176]	Deep Features, Motion Features, and Static Features	42.2 (Untrimmed)
Heilbron et al. 2015 [176]	Deep Features, Motion Features, and Static Features	50.2 (Trimmed)

3.2.3. Discussion

In this section we summarize and discuss the learning-based action representation approaches. These approaches have been divided into genetic programming, dictionary learning and supervised and unsupervised deep learning-based approaches according to the learning representation used in each category. However, this division boundary is not strict and approaches may overlap.

The dictionary learning-based approaches have attracted increasing interest of researchers in computer vision, specifically in human activity recognition. These approaches have introduced the concept of unified learning of dictionary and corresponding classifier into a single learning procedure, which leads to the concept of end-to-end learning. On the other hand, genetic programming (GP) is a powerful evolutionary method inspired by natural selection, used to solve the problem without prior domain knowledge. In human action recognition, GP is used to design the holistic descriptors that are adaptive, and robust for action recognition. These methods have achieved state of the art results on challenging action recognition datasets.

Deep learning has emerged as highly popular direction within the machine learning which has outperformed the traditional approaches in many applications of computer vision. The highly advantageous property of deep learning algorithms is their ability to learn features from the raw data, which eliminates the need of handcrafted feature detectors and descriptors. There are two categories of deep learning models, i.e., unsupervised/generative and supervised/discriminative models. The DBN is a popular generative model which has been used for human action recognition. This model has already achieved high performance on challenging datasets as compared to its traditional handcrafted counterparts [139]. On the other hand, CNN is one of the most popular deep learning models in the supervised learning category. Most of the existing learning-based representations either directly apply CNN to video frames or variations of CNN for spatio-temporal features. These models have also achieved excellent results on challenging human activity recognition datasets as recoded in Table 3. So far, supervised deep learning models have achieved better performance but some studies suggest that unsupervised learning is going to be far more important in the long run. Since we discover the world by observing it rather being told the name of every object, human and animal learning is mostly unsupervised [13].

The deep learning models have also some limitations: These models require huge amount of data for training the algorithm. Most of the action recognition datasets such as KTH [35], IXMAS [118], HDMB-51 [47], and UCF Sports [43,44] are comparatively small for training these models. However, recently a large-scale ActivityNet dataset [176] was proposed with 200 action categories, 849 hours of video in total. This dataset is suitable to train the deep learning-based algorithms. We can expect a major breakthrough with development of algorithms that could produce remarkable results on this dataset.

4. Datasets

In this section well-known public datasets for human activity recognition are discussed. We focus on recently developed datasets which have been frequently used for experimentations.

4.1. Weizmann Human Action Dataset

This dataset [114] was introduced by the Weizmann institute of Science in 2005. This dataset consists of 10 simple actions with static background, i.e., walk, run, skip, jack, jump forward or jump, jump in place or pjump, gallop-sideways or side, bend, wave1, and wave2. It is considered as a good benchmark for evaluation of algorithms proposed for recognition of simple actions. Some methods such as [86,87] have reported 100% accuracy on this dataset. The background of the dataset is simple and only one person performs the action in each frame as shown in Figure 11.

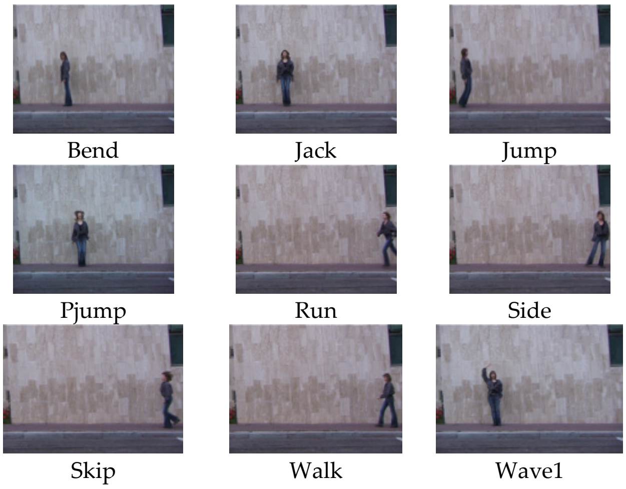


Figure 11. One frame example of each action in Weizmann dataset.

4.2. KTH Human Action Dataset

The KTH dataset [35] was created by the Royal Institute of Technology, Sweden in 2004. This dataset consists of six types of human actions (walking, jogging, running, boxing, hand clapping and hand waving) performed by 25 actors with 4 different scenarios. Thus, it contains 25 × 6 × 4 = 600 video sequences. These videos were recorded with static camera and background; therefore, this dataset is also considered relatively simple for evaluation of human activity recognition algorithms. The method proposed in [36] achieved 98.2% accuracy on this dataset, which is the highest accuracy reported so far. The one frame example of each action from four different scenarios is shown in Figure 12.

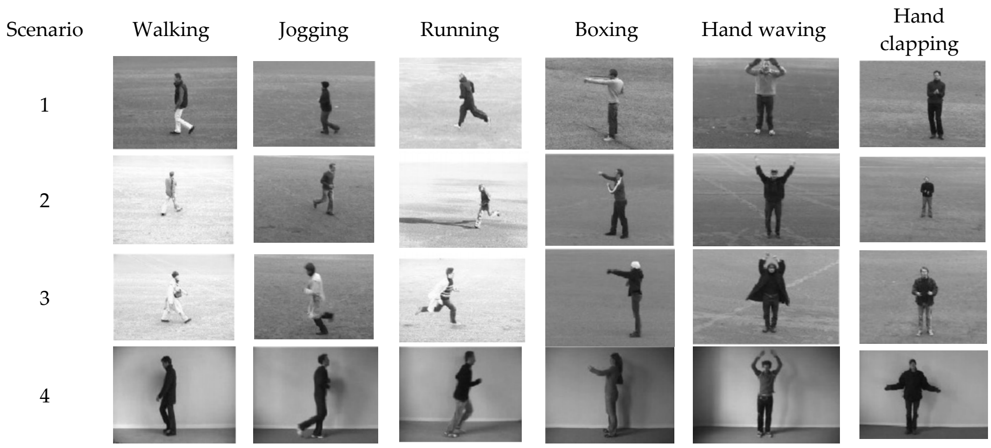


Figure 12. One frame example of each action from four different scenarios in the KTH dataset.

4.3. IXMAS Dataset

INRIA Xmas Motion Acquisition Sequences (IXMAS) [118] a multi-view dataset was developed for evaluation of view-invariant human action recognition algorithms in 2006. This dataset consists of 13 daily life actions performed by 11 actors 3 times each. These actions include crossing arms, stretching head, sitting down, checking watch, getting up, walking, turning around, punching, kicking, waving, picking, pointing, and throwing. These actions were recoded with five calibrated cameras including 4 side cameras and a top camera. The extracted silhouettes of the video sequences are also provided for experimentation. Basically, two types of approaches have been proposed for multi-view action recognition, i.e., 2D and 3D-based approaches. The 3D approaches have reported higher accuracies than the 2D approaches on this dataset but at higher computational cost. The highest accuracy reported on this dataset is 100% in [120] using 3D motion descriptors (HOF3D descriptors and 3D spatial pyramids (SP)). The example frames for each action from five different camera views are shown in Figure 13.

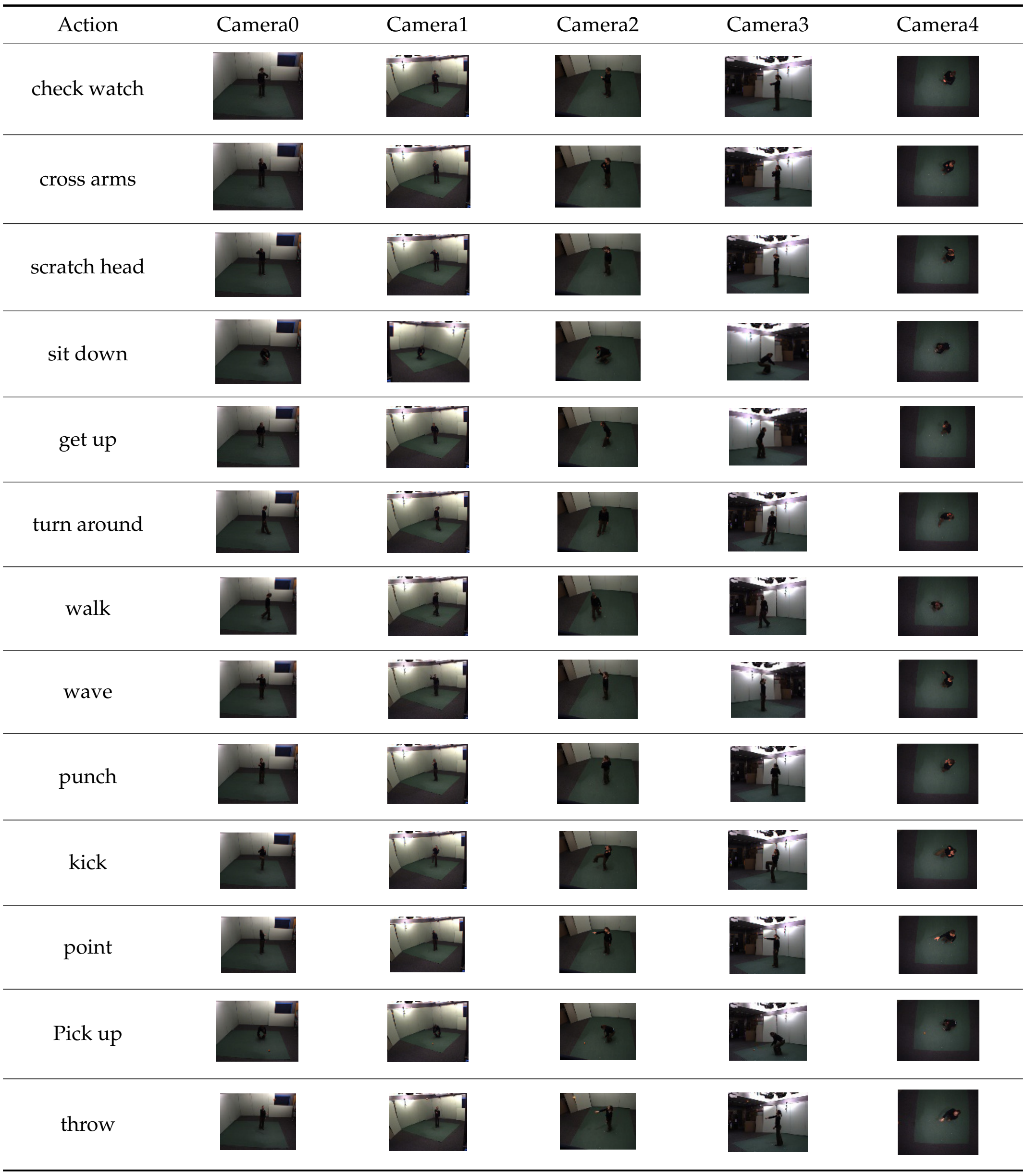


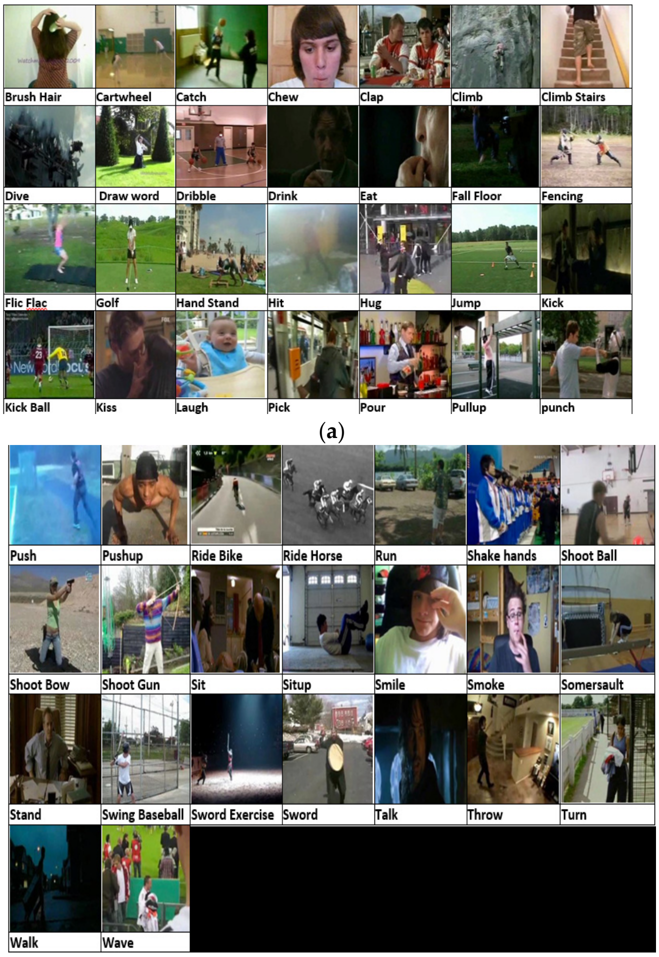
Figure 13. One frame example for each action from five different camera views in IXMAS (INRIA Xmas Motion Acquisition Sequences) dataset.

4.4. HMDB-51

The HMDB-51 [47] is one of the largest datasets available for activity recognition developed by Serre lab, Brown University, USA in 2011. It consists of 51 types of daily life actions comprised of 6849 video clips collected from different sources such as movies, YouTube, and Google videos. The highest accuracy reported so far on this dataset is 74.7% in [157] using SP-CNN (as shown in Table 4). One frame example for each action is shown in Figure 14a,b.

Table 4. Well-known public datasets for human activity recognition.

Dataset	Year	No. of Actions	Method	Highest Accuracy
KTH	2004	6	[36]	98.2%
Weizmann	2005	9	[87]	100%
IXMAS	2006	13	[120]	100%
UCF Sports	2008	10	[36]	95.0%
Hollywood2	2009	12	[169]	75.2%
YouTube	2009	11	[52]	93.38%
HDMB-51	2011	51	[157]	74.7%
UCF-101	2012	101	[157]	91.6
ActivityNet (Untrimmed)	2015	200	[176]	42.2 (baseline)
ActivityNet (Trimmed)	2015	200	[176]	50.2 (baseline)



(**a**)

(**b**)

Figure 14. (**a**) Exemplar frames for action 1–28 from HMDB (Human Motion Database)-51 action dataset; (**b**) Exemplar frames for action 29–51 from HMDB-51 action dataset.

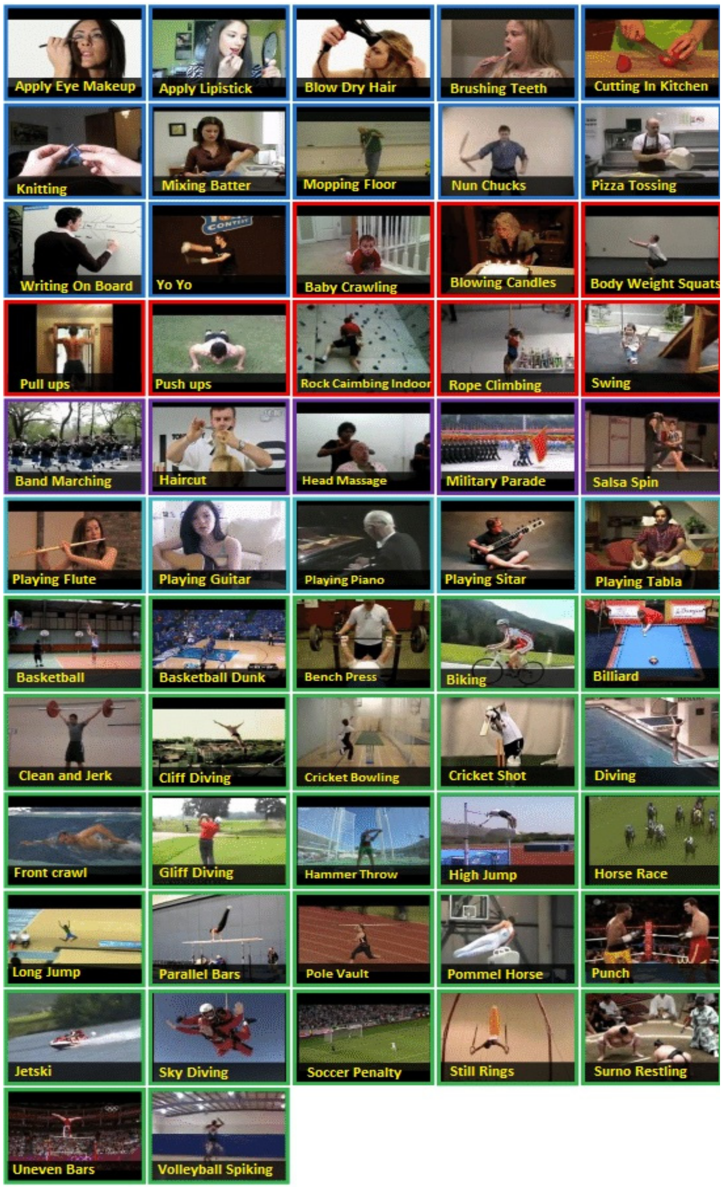
4.5. Hollywood2

Hollywood2 [56] action dataset was created by INRIA (Institut National de Recherche en Informatique et en Automatique), France in 2009. This dataset consists of 12 actions (get out of car, answer phone, kiss, hug, handshake, sit down, stand up, sit up, run, eat, fight, and drive car) with dynamic background features. This dataset is very challenging, consists of short unconstrained movies with multiple persons, cluttered background, camera motion, and large intra-class variations. This dataset is meant for evaluation of HAR algorithms in real life scenarios. Many researchers have evaluated their algorithms on this dataset, the best accuracy achieved so far is 75.2% in [169] using rank pooling and CNN. Some example frames from Hollywood2 dataset are shown in Figure 15.

Figure 15. Exemplar frames from Hollywood2 dataset.

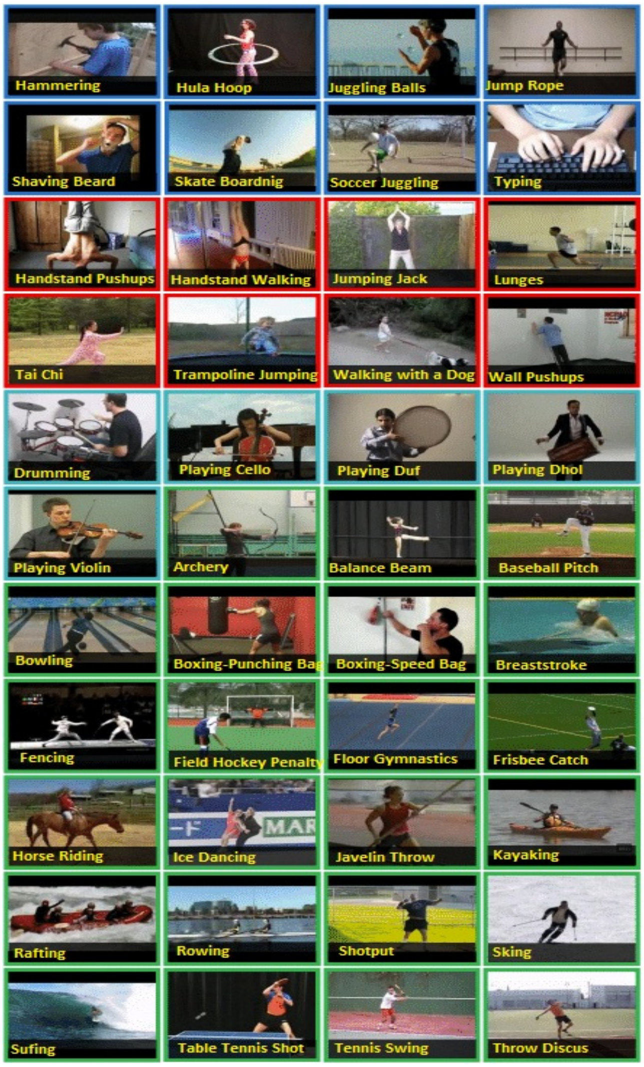
4.6. UCF-101 Action Recognition Dataset

UCF-101 action recognition dataset [173] was created by the Centre for Research in Computer Vision, University of Central Florida, USA in 2012. This is one of the largest action dataset contains 101 action categories collected from YouTube. This dataset is an extension of UCF-50 [177] dataset with 50 action categories. UCF-101 contains 13,320 videos in total, aimed at encouraging the researchers to develop their algorithms for human action recognition is realistic scenarios. The example frames for each action are shown in Figure 16a,b.



(**a**)

Figure 16. *Cont.*



(**b**)

Figure 16. (**a**) Exemplar frames for actions 1–57 from UCF-101 dataset; (**b**) Exemplar frames for actions 58–101 from UCF-101 dataset.

4.7. UCF Sports Action Dataset

UCF sports action dataset was created by the Centre for Research in Computer Vision, University of Central Florida, USA in 2008 [43,44]. It consists of 11 sports action categories (walking, swing-side, swing-bench, skateboarding, running, lifting, kicking, golf swing, riding, and diving) broadcasted on television channels. The dataset includes total 150 video sequences of realistic scenarios. The best accuracy achieved on this dataset so far is 95.0% in [36] using STVs as shown in Table 4. The example frames for each action are shown in Figure 17.

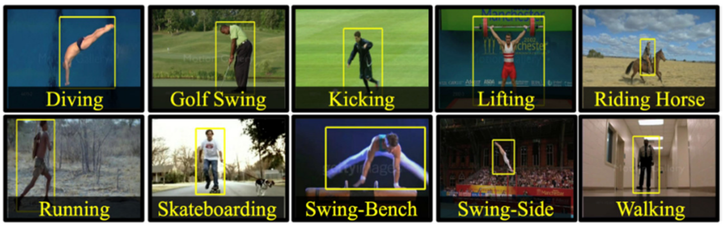


Figure 17. Exemplar frames from sports action dataset.

4.8. YouTube Action Dataset

YouTube action dataset [64] was developed in 2009. This is a challenging dataset due to camera motion, viewpoint variations, illumination conditions, and cluttered backgrounds. It contains 11 action categories: biking, diving, basketball shooting, horse riding, swinging, soccer juggling, trampoline jumping, volleyball spiking, golf swinging, tennis swinging, and walking with a dog. The highest accuracy achieved so far on this dataset is 93.38% in [52] using FV and SFV. The example frames for each action are shown in Figure 18.

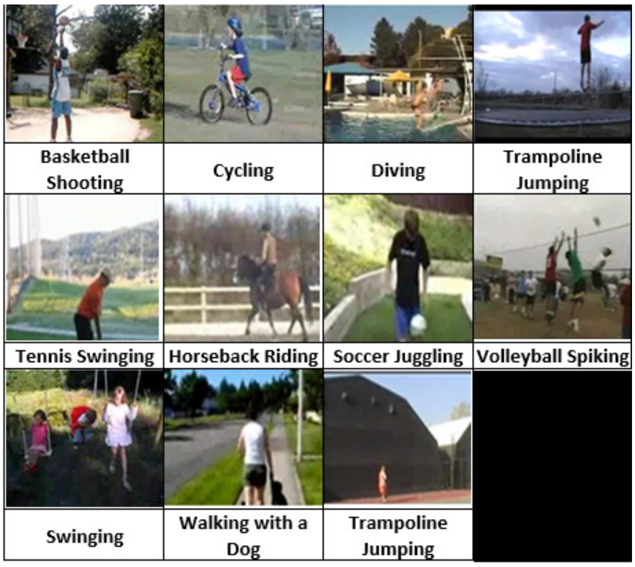


Figure 18. Exemplar frames of 11 sports actions from YouTube action dataset.

4.9. ActivityNet Dataset

ActivityNet [176] was created in 2015. This is a large-scale video dataset covering wide range of complex human activities. It provides 203 action categories in total 849 hours of video data. This dataset is specifically helpful for training the classifiers which require a huge amount of data for training such as deep neural networks. According to the results reported in [176], authors achieved 42.2% accuracy on untrimmed videos and 50.2% on trimmed videos classification. They used deep features (DF), motion features (MF), and static features (SF) as shown in Table 3. Some example frames from this dataset are shown in Figure 19.

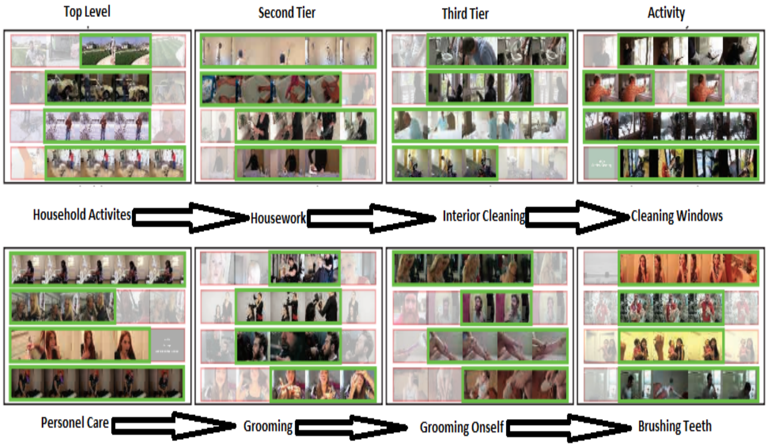


Figure 19. Exemplar frames from ActivityNet dataset.

5. Applications

There are numerous applications of human activity recognition methods. These applications include but are not limited to intelligent video surveillance, entertainment, ambient assisted living, human-robot interaction, and intelligent driving. These applications and state-of-the-art methods and techniques developed for these applications are discussed in the subsequent sections.

5.1. Intelligent Video Surveillance

Traditional security surveillance systems use many cameras and require laborious human monitoring for video content analysis. On the other hand, intelligent video surveillance systems are aimed at automatically tracking the individuals or a crowd and can recognize their activities. These kinds of systems can also detect the suspicious or criminal activities and report to the authorities for immediate action. In this way, the workload of the security personnel can be reduced and security events can be alerted, which can be helpful to prevent dangerous situations.

Different techniques have been proposed for the video surveillance systems. In [178], a visual surveillance system was proposed for tracking and detection of moving objects in real time. A novel method for tracking the pedestrians in crowd video scenes using local spatio-temporal patterns exhibited by the pedestrians was proposed in [179]. It used the collection of Hidden Markov Models (HMMs) trained on local spatio-temporal motion patterns. By employing this representation authors were able to predict the next local spatio-temporal patterns of the tracked pedestrians based on the observed frames of the videos. In [180], a framework was developed for an anomaly detection and automatic behaviour profiling without manual labelling of the training dataset. This framework consists of different components required to develop an effective behaviour representation for event detection.

A real time approach for novelty detection and anomaly recognition in video surveillance systems was proposed in [181]. It is based on on-line clustering technique for the anomaly detection in videos using two-step process. In the first step, recursive density estimation (RDE) was used for novelty detection using frame wise Cauchy type kernel. In the second step, multi-feature on-line clustered trajectory was used for the identification of anomalies in the video stream. The detail survey on video surveillance can be found at [182,183].

5.2. Ambient Assisted Living

HAR-based systems have important applications in the ambient assisted living (AAL). These systems are used as healthcare systems to understand and analyse the patients' activities, to facilitate health workers in treatment, diagnosis, and general healthcare of the patients. In addition to this, these systems are also useful for smart homes such as monitoring daily life activities of elderly people, meant to provide a safe, independent and comfortable stay for elderly people. Usually, these systems capture the continuous movement of elderly people, automatically recognize their activities and detect any abnormality as it happens, such as falling down, having a stroke or respiration issues.

Among these abnormal activities, 'fall' is a major cause for fatal injury, especially for elderly people and is considered a major hindrance for independent living. Different methods have been proposed in the literature for daily life monitoring of elderly people [184]. The fall detection approaches are divided into three types; ambience-device-based, wearable-device-based, and vision-based. Among these, vision-based approaches are more popular and offer multiple advantages as compared to ambience-device-based and wearable-device-based approaches. In [184], a method for fall detection was proposed using combination of integrated time motion image (ITMI) and Eigen space method. The ITMI is a spatio-temporal database which contains motion information and its time stamp and Eigen space technique was used for feature reduction. Then, the reduced feature vector was passed to the neural network for activity recognition. A classification method for fall detection using deformation of the human shape was proposed in [185]. In this method, edge points from the silhouettes were extracted using Canny edge operator for matching two shapes. The Procrustes analysis and mean matching cost were applied for shape analysis. Finally, the fall is characterised by the peak of the smoothed mean matching curve or Procrustes curve. A detailed survey on fall detection approaches can be found in [186].

Monitoring patients' activities remotely is very important for assessing their well-being when they are shifted from hospital to their homes. The homes which are equipped with monitoring facilities are known as smart homes. Different techniques have been proposed in literature for smart homes [187]. Beside, vision-based techniques, a number of sensor-based techniques have also been presented in the literature [188]. In [189], a genetic programming-based classifier was proposed for activity recognition in a smart home environment. It combined the measurement level decisions of the four different classifiers Artificial Neural Network (ANN), Hidden Markov Model (HMM), and Support Vector Machine (SVM) with respect to the assigned weights to each activity class. The weights are optimized using genetic programming for each classifier, where the weights are represented as chromosomes in the form of strings of real values. The results indicate the better performance of the ensemble classifier as compared to a single classifier.

5.3. Human-Robot Interaction

Vision-based activity recognition has important applications in human-robot interaction (HRI). Giving robot the ability to recognize human activities is very important for HRI. This makes the robots useful for the industrial setup and well as in the domestic environment as a personal assistant. In the domestic environment, one of the applications of HRI can be seen as humanoid robots that could recognize human emotions from the sequence of images. A method based on the neural networks was proposed in [190] for the humanoid robot. In this method, six basic emotions (neutral, happiness, sad, fear, disgust and anger) were recognized from the facial expressions, and the topics embedded in the conversation were also identified and analysed. This method is effective for recognizing basic emotions but was not effective for compound emotions such as anger and surprise. In [191], an activity recognition method was proposed for HRI in industrial settings. It was based on the spatial and temporal features from skeletal data of the human workers performing the assembly task. The 3D coordinates of skeletal joints were acquired from Red Green Blue-Depth (RGB-D) data with the help of two Microsoft Kinect sensors. In order to select the best features, the random forests algorithm was applied. Finally, three groups of activities (movement, gestures, and object handling) were recognized using a hidden Markov model. Results indicated an average accuracy of 74.82%. However, this method was unable to recognize more complex activities such as entering and leaving the scene.

Since robots are becoming part of our lives, it is very important that robots should understand human's emotions, intentions, and behaviour. This problem is termed as robot-centric activity recognition. A method for robot-centric activity recognition was proposed in [192] from the videos recorded during HRI. The purpose of this method was recognition of the human activities from the actor's own viewpoint. Unlike conventional third-person activity recognition, the actor (robot) wearing the camera, was involved in ongoing activity. This is not only recognition of activity in real time but also recognizing the activity before its completion, which is really a challenging task. A method for this type of activity recognition was proposed in [193] from RGB-D videos captured by the robot while physically interacting with the humans. The authors used four different descriptors (spatio-temporal points in RGB and depth data, 3D optical flow, and body posture descriptors) as features. With the combination of these descriptors and SVM, authors achieved recognition accuracy of 85.60%.

5.4. Entertainment

Human activity recognition systems are used for recognition of entertainment activities such as dance [114], and sports [194]. In [194], an object-based method was proposed for sports video sequences. In this method, bowling, pitching, golf swing, downhill skiing, and ski jump actions were recognized. The Dynamic Bayesian Networks (DBNs) were employed for recognition of these actions.

The modelling of a player's action in the game has achieved much attention from the sports community in recent years due to its important implications. In addition to this, modelling the behaviour of a player in real-time can be helpful in adapting the change in the game as it occurs. In [195], a method for tracking the behaviour of players during interaction with the game was proposed using an incremental learning technique. The authors used a stream mining change detection technique based on the novelty detection and incremental learning, and performed a set of simulations on UT2004 commercial game.

5.5. Intelligent Driving

Human activity recognition techniques are also employed to assist drivers by providing different cues regarding the state of the driver while driving a vehicle. It has been reported that the secondary tasks performed by the drivers such as answering the phone, sending or receiving text messages, eating or drinking while operating a vehicle cause inattentiveness which may lead to accidents [196,197]. A multi-modal vision frame-work for driver's activity recognition was proposes in [198]. This method extracted the head, eyes, and hand cues to describe the state of the driver. These cues were fused using support vector machine for activity classification. Another technique for driver-activity recognition was proposed in [199]. This method is based on head and eye tracking of the drivers while operating the vehicle.

6. Conclusions

In this review, we provide a compressive survey of state-of-the-art human action representation and recognition approaches including both handcrafted and learning-based representations. The handcrafted action representation approaches have been there for a quite long time. These approaches have achieved remarkable results on different publically available bench mark datasets. However, most successful handcrafted representation methods are based on the local densely-sampled descriptors, which incur a high computational cost. In these approaches, the important features from the sequence of image frames are extracted to build the feature vector using human engineered feature detectors and descriptors. Then, the classification is performed by training a generic classifier. These approaches include space-time, appearance, local binary patterns, and fuzzy logic-based approaches.

On the other hand, learning-based action representation approaches use trainable feature extractors followed by a trainable classifier, which lead to the concept of end-to-end learning or learning from pixel level to action categories identification. This eliminates the need for handcrafted feature detectors and descriptors used for action representation. These approaches include evolutionary (GP-based), dictionary learning, and deep learning-based approaches. Recently, the research community has paid a lot of attention to these approaches. This is mainly due to their high performance as compared to their handcrafted counterparts on some challenging datasets. However, fully data-driven deep models referred to as "black-box" have some limitations: Firstly, it is difficult to incorporate problem-specific prior knowledge into these models. Secondly, some of the best performing deep learning-based methods are still dependent on handcrafted features. The performance of the pure learning-based methods is still not up to the mark. This is mainly due to the unavailability of huge datasets for action recognition unlike in the object recognition where huge dataset such as ImageNet is available. Recently, a large scale dataset ActivityNet has been developed, which is expected to fill this gap. This dataset contains over 200 action categories with 849 hours of video data.

In order to provide further insight into the field, we have presented the well-known public datasets for activity recognition. These datasets include: KTH, Weizmann, IXMAS, UCF Sports, Hollywood2, YouTube, HDMB-51, UCF-101, and ActivityNet. In addition to this, we have also presented the important applications of human activity recognition such as intelligent video surveillance, ambient assisted living, human-robot interaction, entertainment, and intelligent driving.

Acknowledgments: This study is supported by the research grant of COMSATS Institute of Information Technology, Pakistan.

Author Contributions: Allah Bux Sargano, Plamen Angelov, and Zulfiqar Habib conceived and designed the experiments; Allah Bux Sargano performed the experiments; Allah Bux Sargano and Zulfiqar Habib analyzed the data; Allah Bux Sargano and Plamen Angelov contributed reagents/materials/analysis tools; Allah Bux Sargano wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Aggarwal, J.K.; Ryoo, M.S. Human Activity Analysis: A Review. *ACM Comput. Surv. (CSUR)* **2011**, *43*, 16. [CrossRef]
2. Bouwmans, T. Traditional and recent approaches in background modeling for foreground detection: An overview. *Comput. Sci. Rev.* **2014**, *11*, 31–66. [CrossRef]
3. Ke, S.-R.; Thuc, H.L.U.; Lee, Y.-J.; Hwang, J.-N.; Yoo, J.-H.; Choi, K.-H. A review on video-based human activity recognition. *Computers* **2013**, *2*, 88–131. [CrossRef]
4. Ramanathan, M.; Yau, W.-Y.; Teoh, E.K. Human action recognition with video data: Research and evaluation challenges. *IEEE Trans. Hum. Mach. Syst.* **2014**, *44*, 650–663. [CrossRef]
5. Poppe, R. A survey on vision-based human action recognition. *Image Vis. Comput.* **2010**, *28*, 976–990. [CrossRef]
6. Weinland, D.; Ronfard, R.; Boyer, E. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.* **2011**, *115*, 224–241. [CrossRef]
7. Ziaeefard, M.; Bergevin, R. Semantic human activity recognition: A literature review. *Pattern Recognit.* **2015**, *48*, 2329–2345. [CrossRef]
8. Maravelakis, E.; Konstantaras, A.; Kilty, J.; Karapidakis, E.; Katsifarakis, E. Automatic building identification and features extraction from aerial images: Application on the historic 1866 square of Chania Greece. In Proceedings of the 2014 International Symposium on Fundamentals of Electrical Engineering (ISFEE), Bucharest, Romania, 28–29 November 2014.
9. Jalal, A.; Kamal, S.; Kim, D. A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments. *Sensors* **2014**, *14*, 11735–11759. [CrossRef] [PubMed]
10. Jalal, A.; Sarif, N.; Kim, J.T.; Kim, T.S. Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home. *Indoor Built Environ.* **2013**, *22*, 271–279. [CrossRef]
11. Li, J.; Allinson, N. Building recognition using local oriented features. *IEEE Trans. Ind. Inform.* **2013**, *9*, 1697–1704. [CrossRef]
12. Jalal, A.; Kamal, S.; Kim, D. Shape and motion features approach for activity tracking and recognition from kinect video camera. In Proceedings of the 2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops (WAINA), Gwangju, Korea, 25–27 March 2015.
13. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
14. Yuan, R.; Hui, W. Object identification and recognition using multiple contours based moment invariants. In Proceedings of the 2008 International Symposium on Information Science and Engineering, Shanghai, China, 20–22 December 2008.
15. Jalal, A.; Rasheed, Y.A. Collaboration achievement along with performance maintenance in video streaming. In Proceedings of the IEEE Conference on Interactive Computer Aided Learning, Villach, Austria, 26–28 September 2007.
16. Kamal, S.; Azurdia-Meza, C.A.; Lee, K. Subsiding OOB Emission and ICI Power Using iPOWER Pulse in OFDM Systems. *Adv. Electr. Comput. Eng.* **2016**, *16*, 79–86. [CrossRef]
17. Farooq, A.; Jalal, A.; Kamal, S. Dense RGB-D map-based human tracking and activity recognition using skin joints features and self-organizing map. *KSII Trans. Internet Inf. Syst.* **2015**, *9*, 1856–1869.
18. Jalal, A.; Kim, S. The mechanism of edge detection using the block matching criteria for the motion estimation. In Proceedings of the Conference on Human Computer Interaction, Daegu, Korea, 1–4 February 2005.
19. Kamal, S.; Jalal, A. A Hybrid Feature Extraction Approach for Human Detection, Tracking and Activity Recognition Using Depth Sensors. *Arab. J. Sci. Eng.* **2016**, *41*, 1043–1051. [CrossRef]

20. Azurdia-Meza, C.A.; Falchetti, A.; Arrano, H.F. Evaluation of the improved parametric linear combination pulse in digital baseband communication systems. In Proceedings of the 2015 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, 28–30 October 2015.
21. Bongale, P.; Ranjan, A.; Anand, S. Implementation of 3D object recognition and tracking. In Proceedings of the 2012 International Conference on Recent Advances in Computing and Software Systems (RACSS), Chennai, India, 25–27 April 2012.
22. Kamal, S.; Jalal, A.; Kim, D. Depth Images-based Human Detection, Tracking and Activity Recognition Using Spatiotemporal Features and Modified HMM. *J. Electr. Eng. Technol.* **2016**, *11*, 1921–1926.
23. Lai, K.; Bo, L.; Ren, X.; Fox, D. Sparse distance learning for object recognition combining RGB and depth information. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China, 9–13 May 2011.
24. Jalal, A.; Kim, J.T.; Kim, T.-S. Development of a life logging system via depth imaging-based human activity recognition for smart homes. In Proceedings of the International Symposium on Sustainable Healthy Buildings, Seoul, Korea, 19 September 2012.
25. Chang, J.-Y.; Shyu, J.-J.; Cho, C.-W. Fuzzy rule inference based human activity recognition. In Proceedings of the 2009 IEEE Control Applications, (CCA) & Intelligent Control, (ISIC), St. Petersburg, Russia, 8–10 July 2009.
26. Holte, M.B.; Tran, C.; Trivedi, M.M. Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE J. Sel. Top. Signal Process.* **2012**, *6*, 538–552. [CrossRef]
27. Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* **2011**, *2*, 27. [CrossRef]
28. 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. [CrossRef]
29. Sipiran, I.; Bustos, B. Harris 3D: A robust extension of the Harris operator for interest point detection on 3D meshes. *Vis. Comput.* **2011**, *27*, 963–976. [CrossRef]
30. Laptev, I. On space-time interest points. *Int. J. Comput.Vis.* **2005**, *64*, 107–123. [CrossRef]
31. Gilbert, A.; Illingworth, J.; Bowden, R. Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008.
32. Bobick, A.F.; Davis, J.W. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 257–267. [CrossRef]
33. Hu, Y.; Cao, L.; Lv, F.; Yan, S.; Gong, Y. Action detection in complex scenes with spatial and temporal ambiguities. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009.
34. Roh, M.-C.; Shin, H.-K.; Lee, S.-W. View-independent human action recognition with volume motion template on single stereo camera. *Pattern Recognit. Lett.* **2010**, *31*, 639–647. [CrossRef]
35. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, 23–26 August 2004.
36. Sadanand, S.; Corso, J.J. Action bank: A high-level representation of activity in video. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012.
37. Wu, X.; Xu, D.; Duan, L.; Luo, J. Action recognition using context and appearance distribution features. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011.
38. Ikizler, N.; Duygulu, P. Histogram of oriented rectangles: A new pose descriptor for human action recognition. *Image Vis. Comput.* **2009**, *27*, 1515–1526. [CrossRef]
39. Peng, X.; Qiao, Y.; Peng, Q.; Qi, X. Exploring Motion Boundary based Sampling and Spatial-Temporal Context Descriptors for Action Recognition. In Proceedings of the British Machine Vision Conference (BMVC), Bristol, UK, 9–13 September 2013.
40. Liu, J.; Kuipers, B.; Savarese, S. Recognizing human actions by attributes. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011.

41. Chen, M.; Gong, L.; Wang, T.; Feng, Q. Action recognition using lie algebrized gaussians over dense local spatio-temporal features. *Multimed. Tools Appl.* **2015**, *74*, 2127–2142. [CrossRef]
42. Wang, H.; Kläser, A.; Schmid, C. Action recognition by dense trajectories. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011.
43. Rodriguez, M. Spatio-temporal Maximum Average Correlation Height Templates In Action Recognition And Video Summarization. Ph.D. Thesis, University of Central Florida, Orlando, FL, USA, 2010.
44. Soomro, K.; Zamir, A.R. Action recognition in realistic sports videos. In *Computer Vision in Sports*; Springer: Berlin, Germany, 2014; pp. 181–208.
45. Ma, S.; Sigal, L.; Sclaroff, S. Space-time tree ensemble for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015.
46. Wang, C.; Wang, Y.; Yuille, A.L. An approach to pose-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
47. Kuehne, H.; Jhuang, H.; Garrote, E. HMDB: A large video database for human motion recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
48. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 3–6 December 2013.
49. Jiang, Y.-G.; Dai, Q.; Xue, X.; Liu, W.; Ngo, C.W. Trajectory-based modeling of human actions with motion reference points. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012.
50. Kliper-Gross, O.; Gurovich, Y.; Hassner, T. Motion interchange patterns for action recognition in unconstrained videos. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012.
51. Wang, L.; Qiao, Y.; Tang, X. Motionlets: Mid-level 3D parts for human motion recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, Oregon, 25–27 June 2013.
52. Peng, X.; Zou, C.; Qiao, Y.; Peng, Q. Action recognition with stacked fisher vectors. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2014.
53. Jain, M.; Jegou, H.; Bouthemy, P. Better exploiting motion for better action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013.
54. Fernando, B.; Gavves, E.; Oramas, J.M. Modeling video evolution for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015.
55. Hoai, M.; Zisserman, A. Improving human action recognition using score distribution and ranking. In *Asian Conference on Computer Vision*; Springer: Berlin, Germany, 2014.
56. Marszalek, M.; Laptev, I.; Schmid, C. Actions in context. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–26 June 2009.
57. Vig, E.; Dorr, M.; Cox, D. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2012.
58. Mathe, S.; Sminchisescu, C. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *Computer Vision–ECCV 2012*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 842–856.
59. Kihl, O.; Picard, D.; Gosselin, P.-H. Local polynomial space-time descriptors for action classification. *Mach. Vis. Appl.* **2016**, *27*, 351–361. [CrossRef]
60. Lan, T.; Zhu, Y.; Zamir, A.R.; Savarese, S. Action recognition by hierarchical mid-level action elements. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
61. Yuan, J.; Liu, Z.; Wu, Y. Discriminative subvolume search for efficient action detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, Miami, FL, USA, 20–26 June 2009.
62. Amor, B.B.; Su, J.; Srivastava, A. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1–13. [CrossRef] [PubMed]
63. Zanfir, M.; Leordeanu, M.; Sminchisescu, C. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 3–6 December 2013.
64. Liu, J.; Luo, J.; Shah, M. Recognizing realistic actions from videos "in the wild". In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–26 June 2009.

65. Yilmaz, A.; Shah, M. Actions sketch: A novel action representation. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005.
66. Sheikh, Y.; Sheikh, M.; Shah, M. Exploring the space of a human action. In Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV'05), Beijing, China, 17–20 October 2005; Volume 1.
67. Yang, J.; Shi, Z.; Wu, Z. Vision-based action recognition of construction workers using dense trajectories. *Adv. Eng. Inform.* **2016**, *30*, 327–336. [CrossRef]
68. Jiang, Y.-G.; Dai, Q.; Liu, W.; Xue, X. Human Action Recognition in Unconstrained Videos by Explicit Motion Modeling. *IEEE Trans. Image Process.* **2015**, *24*, 3781–3795. [CrossRef] [PubMed]
69. Dollár, P.; Rabaud, V.; Cottrell, G. Behavior recognition via sparse spatio-temporal features. In Proceedings of the 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, 15–16 October 2005.
70. Thi, T.H.; Zhang, J.; Cheng, L.; Wang, L. Human action recognition and localization in video using structured learning of local space-time features. In Proceedings of the 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Boston, MA, USA, 29 August–1 September 2010.
71. Sivic, J.; Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 14–17 October 2003.
72. 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. [CrossRef]
73. Liu, L.; Wang, L.; Liu, X. In defense of soft-assignment coding. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
74. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010.
75. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.L. Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vis.* **2013**, *103*, 60–79. [CrossRef]
76. Li, H.; Greenspan, M. Multi-scale gesture recognition from time-varying contours. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Beijing, China, 17–20 October 2005; Volume 1.
77. Thurau, C.; Hlavác, V. Pose primitive based human action recognition in videos or still images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, Anchorage, AK, USA, 24–26 June 2008.
78. Efros, A.A.; Berg, A.C.; Mori, G. Recognizing action at a distance. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 14–17 October 2003.
79. Fathi, A.; Mori, G. Action recognition by learning mid-level motion features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, Anchorage, AK, USA, 24–26 June 2008.
80. Jiang, Z.; Lin, Z.; Davis, L. Recognizing human actions by learning and matching shape-motion prototype trees. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 533–547. [CrossRef] [PubMed]
81. Holte, M.B.; Moeslund, T.B.; Nikolaidis, N. 3D human action recognition for multi-view camera systems. In Proceedings of the 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), Hangzhou, China, 16–19 May 2011.
82. Huang, P.; Hilton, A.; Starck, J. Shape similarity for 3D video sequences of people. *Int. J. Comput. Vis.* **2010**, *89*, 362–381. [CrossRef]
83. Weinland, D.; Ronfard, R.; Boyer, E. Free viewpoint action recognition using motion history volumes. *Comput. Vis. Image Underst.* **2006**, *104*, 249–257. [CrossRef]
84. Slama, R.; Wannous, H.; Daoudi, M.; Srivastava, A. Accurate 3D action recognition using learning on the Grassmann manifold. *Pattern Recognit.* **2015**, *48*, 556–567. [CrossRef]
85. Wang, L.; Suter, D. Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 18–23 June 2007.
86. Rahman, S.A.; Cho, S.-Y.; Leung, M.K. Recognising human actions by analysing negative spaces. *IET Comput. Vis.* **2012**, *6*, 197–213. [CrossRef]
87. Vishwakarma, D.; Kapoor, R. Hybrid classifier based human activity recognition using the silhouette and cells. *Expert Syst. Appl.* **2015**, *42*, 6957–6965. [CrossRef]

88. Junejo, I.N.; Junejo, K.N.; Al Aghbari, Z. Silhouette-based human action recognition using SAX-Shapes. *Vis. Comput.* **2014**, *30*, 259–269. [CrossRef]
89. Chaaraoui, A.A.; Climent-Pérez, P.; Flórez-Revuelta, F. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognit. Lett.* **2013**, *34*, 1799–1807. [CrossRef]
90. Chaaraoui, A.A.; Flórez-Revuelta, F. A Low-Dimensional Radial Silhouette-Based Feature for Fast Human Action Recognition Fusing Multiple Views. *Int. Sch. Res. Not.* **2014**, *2014*, 547069. [CrossRef] [PubMed]
91. Rahman, S.A.; Song, I.; Song, I.; Leung, M.K.H.; Lee, I. Fast action recognition using negative space features. *Expert Syst. Appl.* **2014**, *41*, 574–587. [CrossRef]
92. Cheema, S.; Eweiwi, A.; Thurau, C. Action recognition by learning discriminative key poses. In Proceedings of the 2011 IEEE. International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011.
93. Chun, S.; Lee, C.-S. Human action recognition using histogram of motion intensity and direction from multiple views. *IET Comput. Vis.* **2016**, *10*, 250–257. [CrossRef]
94. Murtaza, F.; Yousaf, M.H.; Velastin, S. Multi-view Human Action Recognition using 2D Motion Templates based on MHIs and their HOG Description. *IET Comput. Vis.* **2016**, *10*, 758–767. [CrossRef]
95. Ahmad, M.; Lee, S.-W. HMM-based human action recognition using multiview image sequences. In Proceedings of the 18th International Conference on Pattern Recognition, ICPR 2006, Hong Kong, China, 20–24 August 2006.
96. Vishwakarma, D.K.; Kapoor, R.; Dhiman, A. A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics. *Robot. Auton. Syst.* **2016**, *77*, 25–38. [CrossRef]
97. Pehlivan, S.; Forsyth, D.A. Recognizing activities in multiple views with fusion of frame judgments. *Image Vis. Comput.* **2014**, *32*, 237–249. [CrossRef]
98. Eweiwi, A.; Cheema, S.; Thurau, C. Temporal key poses for human action recognition. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011.
99. Ojala, T.; Pietikainen, M.; Harwood, D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In Proceedings of the 12th IAPR International Conference on Pattern Recognition, Jerusalem, Israel, 9–13 October 1994.
100. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
101. Pietikäinen, M.; Hadid, A.; Zhao, G.; Ahonen, T. *Computer Vision Using Local Binary Patterns*; Springer Science & Business Media: London, UK, 2011; Volume 40.
102. Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 915–928. [CrossRef] [PubMed]
103. Yeffet, L.; Wolf, L. Local trinary patterns for human action recognition. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009.
104. Kellokumpu, V.; Zhao, G.; Pietikäinen, M. Human activity recognition using a dynamic texture based method. In Proceedings of the British Machine Vision Conference (BMVC 2008), Leeds, UK, 1–4 September 2008.
105. Kushwaha, A.K.S.; Srivastava, S.; Srivastava, R. Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns. *Multimed. Syst.* **2016**. [CrossRef]
106. Baumann, F.; Ehlers, A.; Rosenhahn, B.; Liao, J. Recognizing human actions using novel space-time volume binary patterns. *Neurocomputing* **2016**, *173*, 54–63. [CrossRef]
107. Sadek, S.; Al-Hamadi, A.; Michaelis, B. An action recognition scheme using fuzzy log-polar histogram and temporal self-similarity. *EURASIP J. Adv. Signal Process.* **2011**, *2011*, 540375. [CrossRef]
108. Yao, B.; Alhaddad, M.J.; Alghazzawi, D. A fuzzy logic-based system for the automation of human behavior recognition using machine vision in intelligent environments. *Soft Comput.* **2015**, *19*, 499–506. [CrossRef]
109. Lim, C.H.; Chan, C.S. Fuzzy qualitative human model for viewpoint identification. *Neural Comput. Appl.* **2016**, *27*, 845–856. [CrossRef]
110. Obo, T.; Loo, C.K.; Seera, M.; Kubota, N. Hybrid evolutionary neuro-fuzzy approach based on mutual adaptation for human gesture recognition. *Appl. Soft Comput.* **2016**, *42*, 377–389. [CrossRef]
111. Yousefi, B.; Loo, C.K. Bio-Inspired Human Action Recognition using Hybrid Max-Product Neuro-Fuzzy Classifier and Quantum-Behaved PSO. *arXiv*, 2015; arXiv:1509.03789.

112. Iglesias, J.A.; Angelov, P.; Ledezma, A. Creating evolving user behavior profiles automatically. *IEEE Trans. Knowl. Data Eng.* **2012**, *24*, 854–867. [CrossRef]
113. Iglesias, J.A.; Angelov, P.; Ledezma, A. Evolving classification of agents' behaviors: A general approach. *Evol. Syst.* **2010**, *1*, 161–171. [CrossRef]
114. 1Gorelick, L.; Blank, M.; Shechtman, E. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2247–2253. [CrossRef] [PubMed]
115. Kellokumpu, V.; Zhao, G.; Pietikäinen, M. Recognition of human actions using texture descriptors. *Mach. Vis. Appl.* **2011**, *22*, 767–780. [CrossRef]
116. Sadek, S.; Al-Hamadi, A.; Michaelis, B. Human action recognition via affine moment invariants. In Proceedings of the 2012 21st International Conference on Pattern Recognition (ICPR), Tsukuba, Japan, 11–15 November 2012.
117. Mattivi, R.; Shao, L. Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor. In *Computer Analysis of Images and Patterns*; Springer: Berlin/Heidelberg, Germany, 2009.
118. Weinland, D.; Boyer, E.; Ronfard, R. Action recognition from arbitrary views using 3D exemplars. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–20 October 2007.
119. Sargano, A.B.; Angelov, P.; Habib, Z. Human Action Recognition from Multiple Views Based on View-Invariant Feature Descriptor Using Support Vector Machines. *Appl. Sci.* **2016**, *10*. [CrossRef]
120. Holte, M.B.; Chakraborty, B.; Gonzalez, J. A local 3-D motion descriptor for multi-view human action recognition from 4-D spatio-temporal interest points. *IEEE J. Sel. Top. Signal Process.* **2012**, *6*, 553–565. [CrossRef]
121. Turaga, P.; Veeraraghavan, A.; Chellappa, R. Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, Anchorage, AK, USA, 24–26 June 2008.
122. Pehlivan, S.; Duygulu, P. A new pose-based representation for recognizing actions from multiple cameras. *Comput. Vis. Image Underst.* **2011**, *115*, 140–151. [CrossRef]
123. Zhu, F.; Shao, L.; Xie, J.; Fang, Y. From handcrafted to learned representations for human action recognition: A survey. *Image Vis. Comput.* **2016**, *55*, 42–52. [CrossRef]
124. Guha, T.; Ward, R.K. Learning sparse representations for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1576–1588. [CrossRef] [PubMed]
125. Wang, H.; Yuan, C.; Hu, W.; Sun, C. Supervised class-specific dictionary learning for sparse modeling in action recognition. *Pattern Recognit.* **2012**, *45*, 3902–3911. [CrossRef]
126. Zheng, J.; Jiang, Z.; Phillips, P.J.; Chellappa, R. Cross-View Action Recognition via a Transferable Dictionary Pair. In Proceedings of the 2012 British Machine Vision Conference, BMVC 2012, Guildford, UK, 3–7 September 2012.
127. Zheng, J.; Jiang, Z.; Chellappa, R. Cross-View Action Recognition via Transferable Dictionary Learning. *IEEE Trans. Image Process.* **2016**, *25*, 2542–2556. [CrossRef] [PubMed]
128. Zhu, F.; Shao, L. Weakly-supervised cross-domain dictionary learning for visual recognition. *Int. J. Comput. Vis.* **2014**, *109*, 42–59. [CrossRef]
129. Zhu, F.; Shao, L. Correspondence-Free Dictionary Learning for Cross-View Action Recognition. In International Conference on Pattern Recognition (ICPR 2014), Stockholm, Sweden, 24–28 August 2014.
130. Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T. Locality-constrained linear coding for image classification. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010.
131. Liu, L.; Shao, L.; Li, X.; Lu, K. Learning spatio-temporal representations for action recognition: A genetic programming approach. *IEEE Trans. Cybern.* **2016**, *46*, 158–170. [CrossRef] [PubMed]
132. Deng, L.; Yu, D. Deep Learning. *Signal Process.* **2014**, *7*, 3–4.
133. Ivakhnenko, A. Polynomial theory of complex systems. *IEEE Trans. Syst. Man Cybern.* **1971**, *1*, 364–378. [CrossRef]
134. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
135. Hinton, G.E.; Osindero, S.; Teh, Y.-W. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef] [PubMed]

136. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
137. Smolensky, P. *Information Processing in Dynamical Systems: Foundations of Harmony Theory*; DTIC Document; University of Colorado Boulder Computer Science Department: Boulder, CO, USA, 1986.
138. Le, Q.V.; Zou, W.Y.; Yeung, S.Y. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011.
139. Foggia, P.; Saggese, A.; Strisciuglio, N. Exploiting the deep learning paradigm for recognizing human actions. In Proceedings of the 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Seoul, Korea, 26–29 August 2014.
140. Hasan, M.; Roy-Chowdhury, A.K. Continuous learning of human activity models using deep nets. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2014.
141. Ballan, L.; Bertini, M.; Del Bimbo, A.; Seidenari, L.; Serra, G. Effective codebooks for human action representation and classification in unconstrained videos. *IEEE Trans. Multimed.* **2012**, *14*, 1234–1245. [CrossRef]
142. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2012.
143. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
144. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2014.
145. Zeiler, M.D.; Taylor, G.W.; Fergus, R. Adaptive deconvolutional networks for mid and high level feature learning. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
146. Karpathy, A.; Li, F.; Johnson, J. CS231n Convolutional Neural Network for Visual Recognition. Available online: http://cs231n.github.io/ (accessed on 10 August 2016).
147. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T. Large-scale video classification with convolutional neural networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014.
148. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2014.
149. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [CrossRef] [PubMed]
150. Wiskott, L.; Sejnowski, T.J. Slow feature analysis: Unsupervised learning of invariances. *Neural Comput.* **2002**, *14*, 715–770. [CrossRef] [PubMed]
151. Zhang, Z.; Tao, D. Slow feature analysis for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 436–450. [CrossRef] [PubMed]
152. Sun, L.; Jia, K.; Chan, T.-H.; Fang, Y.; Wang, G.; Yan, S. DL-SFA: Deeply-learned slow feature analysis for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014.
153. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015.
154. Sun, L.; Jia, K.; Yeung, D.-Y.; Shi, B.E. Human action recognition using factorized spatio-temporal convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
155. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015.
156. Park, E.; Han, X.; Berg, T.L.; Berg, A.C. Combining multiple sources of knowledge in deep CNNs for action recognition. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–9 March 2016.

157. Yu, S.; Cheng, Y.; Su, S.; Cai, G.; Li, S. Stratified pooling based deep convolutional neural networks for human action recognition. *Multimed. Tools Appl.* **2016**, 1–16. [CrossRef]
158. Ijjina, E.P.; Mohan, C.K. Human action recognition based on motion capture information using fuzzy convolution neural networks. In Proceedings of the 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR), Kolkata, India, 4–7 January 2015.
159. Chéron, G.; Laptev, I.; Schmid, C. P-CNN: Pose-based CNN features for action recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
160. Gkioxari, G.; Girshick, R.; Malik, J. Contextual action recognition with R* CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
161. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014.
162. Rahmani, H.; Mian, A. 3D action recognition from novel viewpoints. In Proceedings of the 2016 Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
163. Alfaro, A.; Mery, D.; Soto, A. Action Recognition in Video Using Sparse Coding and Relative Features. *arXiv*, 2016; arXiv:1605.03222.
164. Luo, Y.; Cheong, L.-F.; Tran, A. Actionness-assisted recognition of actions. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
165. Wang, L.; Qiao, Y.; Tang, X. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015.
166. Lan, Z.; Lin, M.; Li, X.; Hauptmann, A.G.; Raj, B. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015.
167. Bilen, H.; Fernando, B.; Gavves, E.; Vedaldi, A.; Gould, S. Dynamic image networks for action recognition. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition CVPR, Las Vegas, NV, USA, 27–30 June 2016.
168. Mahasseni, B.; Todorovic, S. Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition. In Proceedigs of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 27–30 June 2016.
169. Fernando, B.; Gavves, E.; Oramas, J.; Ghodrati, A.; Tuytelaars, T. Rank pooling for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**. [CrossRef]
170. Zhu, W.; Hu, J.; Sun, G.; Cao, X.; Qiao, Y. A key volume mining deep framework for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
171. Wang, C.; Wang, Y.; Yuille, A.L. Mining 3D key-pose-motifs for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
172. Veeriah, V.; Zhuang, N.; Qi, G.-J. Differential recurrent neural networks for action recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
173. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv*, 2012; arXiv:1212.0402.
174. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015.
175. Weinzaepfel, P.; Harchaoui, Z.; Schmid, C. Learning to track for spatio-temporal action localization. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
176. Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; Carlos Niebles, J. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–10 June 2015.
177. Reddy, K.K.; Shah, M. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.* **2013**, *24*, 971–981. [CrossRef]

178. Lizhong, L.; Zhiguo, L.; Yubin, Z. Research on Detection and Tracking of Moving Target in Intelligent Video Surveillance. In Proceedings of the 2012 International Conference on Computer Science and Electronics Engineering (ICCSEE), Hangzhou, China, 23–25 March 2012.
179. Kratz, L.; Nishino, K. Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 987–1002. [CrossRef] [PubMed]
180. Xiang, T.; Gong, S. Video behavior profiling for anomaly detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 893–908. [CrossRef] [PubMed]
181. Sadeghi-Tehran, P.; Angelov, P. A real-time approach for novelty detection and trajectories analysis for anomaly recognition in video surveillance systems. In Proceedings of the 2012 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), Madrid, Spain, 17–18 May 2012.
182. Hu, W.; Tan, T.; Wang, L.; Maybank, S. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* **2004**, *34*, 334–352. [CrossRef]
183. Paul, M.; Haque, S.M.; Chakraborty, S. Human detection in surveillance videos and its applications—A review. *EURASIP J. Adv. Signal Process.* **2013**, *2013*, 176. [CrossRef]
184. Foroughi, H.; Naseri, A.; Saberi, A.; Yazdi, H.S. An eigenspace-based approach for human fall detection using integrated time motion image and neural network. In Proceedings of the 9th International Conference on Signal Processing, ICSP 2008, Leipzig, Germany, 10–11 May 2008.
185. Rougier, C.; Meunier, J.; St-Arnaud, A.; Rousseau, J. Robust video surveillance for fall detection based on human shape deformation. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 611–622. [CrossRef]
186. Mubashir, M.; Shao, L.; Seed, L. A survey on fall detection: Principles and approaches. *Neurocomputing* **2013**, *100*, 144–152. [CrossRef]
187. Benmansour, A.; Bouchachia, A.; Feham, M. Multioccupant activity recognition in pervasive smart home environments. *ACM Comput. Surv. (CSUR)* **2016**, *48*, 34. [CrossRef]
188. Jurek, A.; Nugent, C.; Bi, Y.; Wu, S. Clustering-based ensemble learning for activity recognition in smart homes. *Sensors* **2014**, *14*, 12285–12304. [CrossRef] [PubMed]
189. Fatima, I.; Fahim, M.; Lee, Y.-K.; Lee, S. Classifier ensemble optimization for human activity recognition in smart homes. In Proceedings of the 7th International Conference on Ubiquitous Information Management and Communication, Kota Kinabalu, Malaysia, 17–19 January 2013.
190. Zhang, L.; Jiangb, M.; Faridc, D.; Hossaina, M.A. Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot. *Expert Syst. Appl.* **2013**, *40*, 5160–5168. [CrossRef]
191. Roitberg, A.; Perzylo, A.; Somani, N.; Giuliani, M.; Rickert, M.; Knoll, A. Human activity recognition in the context of industrial human-robot interaction. In Proceedings of the 2014 Annual Summit and Conference (APSIPA) Asia-Pacific Signal and Information Processing Association, Chiang Mai, Thailand, 9–12 December 2014.
192. Ryoo, M.; Fuchs, T.J.; Xia, L.; Aggarwal, J.K.; Matthies, L. Robot-Centric Activity Prediction from First-Person Videos: What Will They Do to Me. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, Portland, OR, USA, 2–5 March 2015.
193. Xia, L.; Gori, I.; Aggarwal, J.K.; Ryoo, M.S. Robot-centric Activity Recognition from First-Person RGB-D Videos. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Beach, HI, USA, 6–9 January 2015.
194. Luo, Y.; Wu, T.-D.; Hwang, J.-N. Object-based analysis and interpretation of human motion in sports video sequences by dynamic Bayesian networks. *Comput. Vis. Image Underst.* **2003**, *92*, 196–216. [CrossRef]
195. Vallim, R.M.; Filho, J.A.A.; De Mello, R.F.; De Carvalho, A.C.P.L.F. Online behavior change detection in computer games. *Expert Syst. Appl.* **2013**, *40*, 6258–6265. [CrossRef]
196. Klauer, S.G.; Guo, F.; Sudweeks, J.; Dingus, T.A. *An Analysis of Driver Inattention Using a Case-Crossover Approach on 100-Car Data: Final Report*; National Highway Traffic Safety Administration: Washington, DC, USA, 2010.
197. Tison, J.; Chaudhary, N.; Cosgrove, L. *National Phone Survey on Distracted Driving Attitudes and Behaviors*; National Highway Traffic Safety Administration: Washington, DC, USA, 2011.

198. Eshed Ohn-Bar, S.M.; Tawari, A.; Trivedi, M. Head, eye, and hand patterns for driver activity recognition. In Proceeedings of the 2014 IEEE International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014.
199. Braunagel, C.; Kasneci, E.; Stolzmann, W.; Rosenstiel, W. Driver-activity recognition in the context of conditionally autonomous driving. In Proceeedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Canary Islands, Spain, 15–18 September 2015.