sensors

MDPI

Review

Video Activity Recognition: State-of-the-Art

Itsaso Rodríguez-Moreno [1,*], José María Martínez-Otzeta [1], Basilio Sierra [1], Igor Rodriguez [1] and Ekaitz Jauregi [2]

[1] Department of Computer Science and Artificial Intelligence, University of the Basque Country, Manuel Lardizabal 1, 20018 Donostia-San Sebastián, Spain

[2] Department of Computer Languages and Systems, University of the Basque Country, Manuel Lardizabal 1, 20018 Donostia-San Sebastián, Spain

* Correspondence: itsaso.rodriguez@ehu.eus; Tel.: +34-943-015-107

Received: 13 June 2019; Accepted: 9 July 2019; Published: 18 July 2019

Abstract: Video activity recognition, although being an emerging task, has been the subject of important research efforts due to the importance of its everyday applications. Surveillance by video cameras could benefit greatly by advances in this field. In the area of robotics, the tasks of autonomous navigation or social interaction could also take advantage of the knowledge extracted from live video recording. The aim of this paper is to survey the state-of-the-art techniques for video activity recognition while at the same time mentioning other techniques used for the same task that the research community has known for several years. For each of the analyzed methods, its contribution over previous works and the proposed approach performance are discussed.

Keywords: activity recognition; computer vision; optical flow; deep learning

1. Introduction

Activity recognition consists of identifying some actions from a series of observations. This field has caught the interest of many researchers since the 1980s due to the number of applications for which it is useful, such as medicine [1,2], human–computer interaction [3,4], surveillance [5,6] or sociology [7,8]. For instance, in surveillance [9,10], the automatic detection of suspicious actions would allow for launching a warning and taking measures against any danger. Another example is the use of activity recognition for rehabilitation [11], recognizing the action the patients are performing and having the ability to determine if it is right or not. One of the main techniques used for activity recognition is computer vision, namely video-based activity recognition. Visual video features provide basic information for video events or actions.

The task of tracking and understanding what is happening in a video can be very challenging. Many attempts have been made lately using different techniques [12–14] such as optical flow [15,16], Hidden Markov Models (HMM) [17–19] or, more recently, deep learning [20,21]. Furthermore, apart from using multiple techniques, many different scenarios are being considered, single action recognition [22,23], group tracking [24,25], etc.

However, despite remarkable progress, the advances achieved so far do not meet high accuracy standards and the correct realization of this task in some areas, such as video surveillance, is still an open research issue.

In the analysis of a video content, many different functionalities can be implemented. One of the simplest ways to detect motion regarding a fixed background is Video Motion Detection [26–28]. Video tracking [29,30] is more challenging than the previous approach and can be very time consuming, due to the amount of data that a video contains. The aim of video tracking is to associate target objects in consecutive video frames, which can be especially difficult if the objects are moving fast in relation to the frame rate. If object recognition techniques are needed (a challenging problem in its own), further complexity is added. On the contrary, the human brain seems to have the ability to recognize human actions perfectly. This aptitude is not just related to acquired knowledge, but also to logical reasoning and the capability of extracting relevant information from context. Based on this, the integration of commonsense reasoning [31,32] and contextual knowledge [33] has been proposed.

Hence, action recognition involves the classification of different actions from videos, a sequence of frames, taking into account as well the fact that the action could not be performed during the entire video. Although it seems an extension of image classification tasks, as it has been mentioned before, the progress for video classification has been slower due to various reasons:

- Apart from spatial information, temporal context across frames is also required.
- Huge computational cost.
- Datasets are more limited, due to the difficulty to collect, annotate and store videos.

Throughout this paper, several techniques applied for video activity recognition are mentioned, as well as the latest contributions made in the field. In addition, as a final note, some of the databases used for this topic are presented along with the results of the latest contributions using them. In Figure 1, a diagram showing the techniques explained and other tasks related to this subject but which are not discussed in this review are indicated.

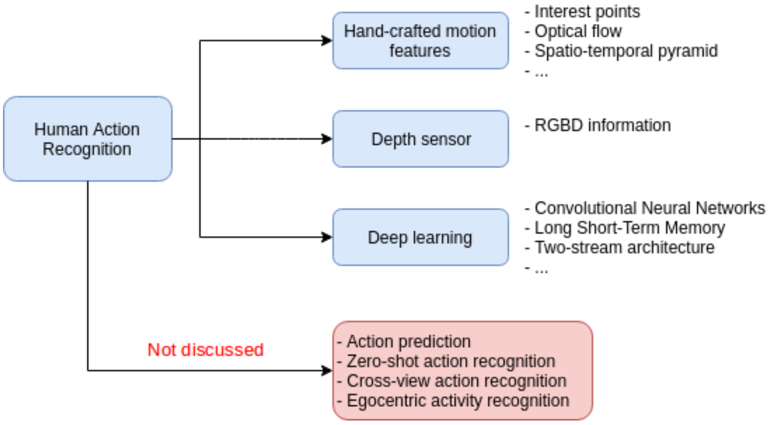
Human Action Recognition
Hand-crafted motion features
- Interest points
- Optical flow
- Spatio-temporal pyramid
- ...
Depth sensor
- RGBD information
Deep learning
- Convolutional Neural Networks
- Long Short-Term Memory
- Two-stream architecture
- ...
Not discussed
- Action prediction
- Zero-shot action recognition
- Cross-view action recognition
- Egocentric activity recognition

Figure 1. Summary diagram.

This review focuses on a specific area of Human Action Recognition, to keep the discussion simple. Only action recognition from a whole video recorded from a fixed position is considered in this paper, as we think this problem setup is the entrance gate to the analysis of other more complex situations, as those presented in the bottom part of Figure 1. At the same time, the complexity level of the problem considered in this review is high enough to deserve a dedicated survey. For the sake of completeness, we will briefly review the main characteristics of the situations shown in Figure 1 but not covered here. In action prediction, instead of recognizing the action that is happening in the video, the objective is to guess the action that will occur in an incomplete video. The zero-shot action recognition problem consists of training a model to classify videos of categories that have no instances in the training set, which means that there are no instances of certain classes that are going to appear in the test set. To address this issue, complementary information of invisible classes is assumed in the form of attribute vectors that describe each class. In the cross-view action recognition, there are different points of view in the scene when the action is occurring. There are other variations such as egocentric activity recognition that consists of recognizing actions from egocentric videos [34].

The survey is centered in action recognition methods for videos that are recorded in third person and the whole action occurs inside the video. Although different information can be extracted from the videos and there are articles mentioned that also use extra information such as depth sensors' information, all the presented methods have these two characteristics in common. The methods that are explained use databases with the characteristics of the ones presented in Section 3. Although there are previous reviews on video action recognition [12,35,36], as it is a subject that is continuously progressing, it is always necessary to have a survey that collects the latest contributions. Our review, apart from mentioning articles that others have not been able to collect since they have been published later, also deals with older articles that have served as reference for later methods.

2. Used Techniques

As activity recognition has been an active research area lately, there have been many different approaches to deal with this problem. Throughout the survey, some of these are introduced, starting with simpler approaches and finishing with the newest contributions to the field. The proposed methods that try to solve this problem that are referred to in this paper could be separated into three main groups: methods using hand-crafted motion features, depth information based methods and deep learning based methods. Strictly speaking, these three areas are somehow interrelated and depth sensors features could lie under the hand-crafted or deep learning categorization. For a long time, computer vision has focused on data recorded from RGB (visible light) cameras, especially in the case of videos. Depth sensors have started to be used in the field of video analysis in more recent times and this is the reason why we feel it deserves a separate section.

First, hand-crafted motion features methods are explained. In these methods, some interesting features are obtained from the raw pixels of the video frames and then these features are used to perform the recognition. Second, depth information based methods are analyzed, which use depth maps as extra information. Third, deep learning methods are presented, which, unlike hand-crafted methods, achieve the features for the recognition automatically. Throughout the document, several methods that combine some of these three modalities are also presented.

2.1. Methods Using Hand-Crafted Motion Features

This document focuses on video-based activity recognition, in which the representation of visual and temporal information becomes important. There are several ways to extract visual features, both static image features and temporal visual features, and then use them to perform the recognition. Temporal visual features are a combination of static image features and time information, so, through these features, temporal video information is achieved. Key-frame [37,38], bag-of-words (BoW) [39,40], interest points [41,42] and motion based approaches [43–46] are types of representations that can be obtained from a video. *Key-frame* based approaches, as the name indicates, consist of detecting the key-frames of the video which would be used for classification; *BoW* based approaches represent the frames of the video segments over a vocabulary of visual features; *interest points* based approaches focus on simply selecting a specific set of points or pixels for the classification and, to finish, *motion* based approaches focus on the movement along the video. Throughout this section, only motion based approaches are analyzed.

In [47], the authors use a temporal template as the basis of their representation, continuing with their approach presented in [48]. This temporal template consists of a static vector-image where the value of the vector at each point represents a function of the motion properties at the corresponding spatial location in an image sequence. They explore their representation with a simple two component version of the template:

- The first value indicates the presence of motion and where it occurs by a binary motion-energy image (MEI). Being $D(x,y,t)$ a binary image sequence and r the value that defines the temporal extent of a movement, the binary image is defined this way:

$$E_r(x,y,t) = \bigcup_{i=0}^{r-1} D(x,y,t-i). \tag{1}$$

- The second value is a scalar-valued image where intensity is a function of recency of motion of the sequence, represented by a motion-history image (MHI) which indicates how the image is moving. H_r represents the temporal history of motion at each point, where recently moved pixels are brighter:
$$H_r(x,y,t) = \begin{cases} r, & \text{if } D(x,y,t) = 1, \\ max(0, H_r(x,y,t-1)-1), & \text{otherwise.} \end{cases} \tag{2}$$

Then, a recognition method is developed, which matches these temporal templates against stored instances of known actions. They also present a recognition method to automatically perform temporal segmentation being invariant to linear changes in speed.

The authors of [49] demonstrate that local measurements in terms of spatio-temporal interest points (local features) can be used to recognize complex motion patterns. As these features, which capture local motion events in videos, can be adapted to size, frequency and velocity of moving patterns, the resulting video representations are stable with respect to the corresponding transformations. To represent motion patterns, they use local space-time features [50] and to detect local features they construct, using Gaussian convolution, its scale-space representation. Then, they explore the integration of local space-time features with Support Vector Machines (SVM) classifier [51,52], used in many visual pattern recognition methods [53,54], and apply the resulting approach to the recognition of human actions. In addition, for the purpose of evaluation, the authors introduce a new video database containing 2391 clips of six human-actions performed by 25 people in four scenarios.

In [55], the authors present a hybrid hierarchical model, inspired by [56], where video sequences are represented as collections of spatial and spatio-temporal features. These features are achieved by extracting both static and dynamic interest points and the model is able to combine static and motion image features, as well as performing categorization of human actions in a frame-by-frame basis. Motion features are extracted as in [40]. They show that using static and dynamic features together is better than using just a single feature type.

Laptev et al. [42] contribute to the recognition of realistic videos and use movie scripts for automatic annotation of human actions in videos. Due to the achievements in image classification [57–60], they employ spatio-temporal features and spatio-temporal pyramids, extending spatial pyramids of [58]. Interests points are detected as in [50] using a space-time extension of the Harris operator [61]. Then, a multi-scale approach is used and features at multiple levels are extracted. For classification, they use a nonlinear SVM with a multi-channel Gaussian kernel [60]. Apart from the action recognition task, their main contribution consists of automatically annotating human actions with the use of movie scripts and getting videos with more realistic characteristics.

Visual features such as edges, corners, interest points, etc. can be used to form a more complicated feature called optical flow. The optical flow methods try to calculate the motion between two image frames which are taken at times t and $t + \Delta t$ at every position, assuming that the intensity of objects does not change during the movement $I(x,y,t) = I(x + \Delta x, y + \Delta y, t + \Delta t)$. Expanding that equation using the Taylor Series Expansion [62] and further calculations, this equation is obtained:

$$I_x V_x + I_y V_y = -I_t \tag{3}$$

or

$$\nabla I^T \cdot \vec{V} = -I_t. \tag{4}$$

The solution, the optical flow, is the value of $\vec{V}$. Some approaches are given in the calculation of optical flow due to the fact that there are two unknowns in the equation. In this part, several methods that have made use of this feature and its variations are presented.

The authors of [63] present a method to recognize human actions observing them from a far field of view, but they also test their model with normal resolution datasets, such as Weizmann [64]. They use Histograms of Oriented Gradients (HOG) for human pose representations, first introduced in [65] and successfully applied in multiple action recognition methods [66–69]. They also use a time series of Histogram of Oriented Optical Flow (HOOF) to characterize human motion. To get a subset of discriminantly informative principal components (PCs), an extension of Supervised Principal Component Analysis (SPCA) [70] technique is used, which tries to select a subset of PCs in order to best separate samples projected from different classes. This step significantly speeds up the run-time of recognition without sacrificing accuracy. A multi-class Support Vector Machine (SVM) classifier is trained for action classification. The classifier prediction is made by a collection of one-against-one SVM classifiers, as in the implementation of [71].

In [46], inspired by the success of histograms of features in object recognition, the authors propose the representation of each frame with the use of HOOF features, which are independent from the scale of the moving person and to the direction of motion. These histograms are created by computing optical flow at every frame and binning the vectors according to each primary angle. To classify HOOF time-series, they posit a generalization of the Binet–Cauchy kernels [72] to nonlinear dynamical systems (NLDS), as the data that represents, for instance, that the histogram time series is non-Euclidean and needs to be modeled with nonlinear dynamical systems. The generalization is done by using a Mercer kernel [73] on the output space. The Binet–Cauchy kernels are used for NDLS to perform the activity recognition and proposed HOOF features as outputs of NLDS.

The authors of [45] introduce a motion descriptor based on direction of optical flow. In their method, interest silhouettes are subtracted from the background (used dataset provides foreground masks [64]) and optical flow is computed using the Lucas–Kanade algorithm [74]. Then, before computing a direction histogram, the window is divided into eight regions. To represent the distribution of optical flow direction, they use a histogram, segmenting the direction of optical flow into eight bins. To create the motion vector, they concatenate a direction histogram of optical flow in every region. They also smooth the motion vectors to reduce motion variation and noise, and then these vectors are used for classification. K-means clustering [75] is first used to group similar postures and then the classification is done by a K-NN classifier. Niebles et al. [55] also used clustering but with a bag-of-words model instead of motion.

Due to the demonstration of dense trajectories being efficient video representations, in [76], their performance is improved by using camera motion to correct them. The estimation of camera motion is done by matching feature points between frames using SURF (speeded up robust features) descriptors [77] and dense optical flow [78]. A human detector [79,80] is used to remove inconsistent matches generated because of the differences of human and camera motions and, in addition, background trajectories are also removed. Motion-based descriptors, such as HOF (histogram optical flow), are significantly improved by this.

In [81], the authors propose a generic temporal video segment representation method for action recognition based on optical flow concept [62], with the idea that, to deal with a video-based action recognition problem, temporally represented video information is needed. In their approach, for feature detection, the Shi–Tomasi algorithm is used [82], which is based on Harris corner detector [61], and, to estimate optical flow, the Lucas–Kanade algorithm [74] is computed. For each selected frame of the video, optical flow vectors are grouped according to their angular features. Being an optical flow histogram the most common method of optical flow based video representation, they enrich these approaches by a novel velocity concept, Weighted Frame Velocity. This concept refers to the velocity of cumulative angular grouping of a temporal video segment, which represents the motion of the frames more descriptively. Similarities in the histogram do not always mean that there are similarities in the motion, so, instead of using a histogram based approach as in [46,83–85], vectors are grouped with respect to their angular characteristics and then summed and integrated with the new velocity concept.

The authors of [15] propose a local descriptor built by optical flow vectors along the edges of the action performers. First, a foreground extraction is done by a Gaussian Mixture Model (GMM) based method [86] and optical flow based technique [62] in order to segment the region of interest. To represent the segmented objects, optical flow based feature vectors are computed along the boundary using Horn and Schunck algorithm [62] based optical flow extraction technique. This way, shape and instantaneous velocity information extracted from the boundaries of the action performers are incorporated in the feature set. These features are then used to feed a multi-class SVM classifier.

In [87], human activities are recognized using background subtraction, HOG features and Back-Propagation Neural Network (BPNN) classifier. In this approach, background estimation is performed at first, using mean filter to obtain the background and areas of the image containing important information. Afterwards, in order to extract features to describe human motion, a histogram of oriented gradients (HOG) [65] descriptor is used, with the idea that local shape information can be completely described by intensity gradients or edge directions. Finally, a BPNN is used to perform the final classification.

In Table 1, a summary of the explained methods using hand-crafted motion features is presented.

Table 1. Summary of methods using hand-crafted motion features.

	YEAR	SUMMARY	DATASET
Bobick et al. [47]	2001	Use of motion-energy image (MEI) and motion-history image (MHI).	-
Schuldt et al. [49]	2004	Use of local space-time features to recognize complex motion patterns.	KTH Action [49]
Niebles et al. [55]	2007	Use of a hybrid hierarchical model, combining static and dynamic features.	Weizmann [64]
Laptev et al. [42]	2008	Use of spatio-temporal features and extend spatial pyramids to spatio-temporal pyramids.	KTH Action [49] Hollywood [42]
Chen et al. [63]	2009	Use of HOG for human pose representations and HOOF to characterize human motion.	Weizmann [64] Soccer [83] Tower [63]
Chaudhry et al. [46]	2009	Use of HOOF features by computing optical flow at every frame and binning them according to primary angles.	Weizmann [64]
Lertniphonphan et al. [45]	2011	Use of a motion descriptor based on direction of optical flow.	Weizmann [64]
Wang et al. [76]	2013	Use of camera motion to correct dense trajectories.	HMDB51 [88] UCF101 [89] Hollywood2 [90] Olympic Sports [91]
Akpinar et al. [81]	2014	Use of a generic temporal video segment representation, introducing a new velocity concept: Weighted Frame Velocity.	Weizmann [64] Hollywood [42]
Kumar et al. [15]	2016	Use of a local descriptor built by optical flow vectors along the edges of the action performers.	Weizmann [64] KTH Action [49]
Sehgal, S. [87]	2018	Use of background subtraction, HOG features and BPNN classifier.	Weizmann [64]

2.2. Depth Information Based Methods

The interest of applying depth data captured from depth cameras for the action recognition problem has grown due to the advances of imaging technology in capturing depth information in real time, such as Microsoft Kinect [92] and Intel Realsense [93]. In the past few decades, research of human action recognition has mainly concentrated on video sequences captured by traditional RGB cameras, but, thanks to the advances in imaging techniques, RGBD sensors are able to capture color image sequences together with depth maps in real time. Depth images are insensitive to changes in lighting conditions and provide additional body shape and motion information that can help with distinguishing actions that generate similar projections from a single view. In this paper, some of the recent methods using depth maps are introduced. However, if more information is required, there are many other interesting methods to analyze [94–97].

In [98], the authors propose the use of sequences of depth maps for action recognition, which provide additional body shape and motion information. In their approach, in order to make use of the additional body shape and motion information from depth maps, they generate Depth Motion Maps (DMM) by projecting depth maps into three ortoghonal Cartesian planes and accumulating global activities through entire video sequences. Then, a good characterization of the local appearance and shape on DMM is achieved with HOG, Histrogram of Oriented Gradients. HOG descriptors extracted from depth motion map of each projection view (front, top, side) are combined as DMM-HOG, which is used to represent the entire action video sequences. This DMM-HOG descriptor is the input to a linear SVM classifier which is used to make the recognition.

A new descriptor for activity recognition from videos obtained with a depth sensor is presented in [99], called the histogram of oriented 4D surface normals (HON4D). In order to capture the complex joint shape-motion cues at the pixel level, the authors use a histogram to describe depth sequence, which captures the distribution of the surface normal orientation in 4D space of time, depth and spatial coordinates. Instead of concatenating features [100], their histogram, as it operates in 4D space, captures the distribution of the changing shape and motion cues along with their correlation. The histogram is built by creating 4D projectors that represent the possible directions of the 4D normal and, as the descriptor is a representation for the entire sequence, it is robust against noise and occlusion, unlike other methods [101]. To quantize the 4D space, they use the vertices of a polychoron to get a more discriminative quantification.

In [102], the authors present a two-layer Bag-of-Visual-Words (BoVW) model. First, they delete background clutter, so background noise is removed. In addition, foreground noise disturbances are eliminated by jointly using motion and shape information. To distinguish similar actions, motion-based STIPs (spatial-temporal interest points) and shape based STIPs are detected. They use 3DLSK, first mentioned in [103], to describe local structures of motion-based STIPs, and, in order to fit better to depth data and its lack of texture or scale changes effects, they propose a multi-scale 3DLSK (M3DLSK). On the other hand, to capture spatial-temporal relationships among STIPs, they extract a spatial-temporal vector (STV) descriptor for each STIP to distinguish between different actions. Fusing both descriptors, M3DLSK and STV, a feature representation able to capture local and global motion and shape is achieved.

Satyamurthi et al. [104] propose the use of depth motion maps projected on multiple directions, multi-directional projected depth motion map (MPDMM), based on depth motion maps [96,98]. The proposed approach can be separated in three key components. First, they propose to extract features by converting the video sequences into frames using multi-directional projected DMM. The input 3D depth action video is projected into a set of 2D maps according to a set of planes and directions. After calculating the motion energy of each projected map, this is concatenated through entire video sequences to get the MPDMM model. Second, features are extracted from MPDMM model, on the basis of conventional texture-based Local Binary Patterns (LBP) descriptors [105]. The MPDMM image is processed with the LBP technique by thresholding the neighborhood of each pixel and outputting the result as a series of binary numbers that are then used as a statistical measure forming a histogram. Third, the kernel-based Extreme Learning Machine (ELM) [106] with a radial basis function kernel is applied to perform the classification.

In Table 2, a summary of the explained depth information based methods is presented.

Table 2. Summary of depth information based methods.

	YEAR	SUMMARY	DATASET
Yang et al. [98]	2012	Use of Depth Motion Maps (DMM), combining them with HOG descriptors.	MSRAction3D [107]
Oreifej et al. [99]	2013	Use of histogram of oriented 4D surface normals (HON4D) descriptor.	MSRAction3D [107] MSRGesture3D [108] 3D Action Pairs [99]
Liu et al. [102]	2018	Use of a two-layer BoVW model, using motion-based and shape-based STIPs to distinguish the action.	MSRAction3D [107] UTKinect-Action [109] MSRGesture3D [108] MSRDailyActivity3D [100]
Satyamurthi et al. [104]	2018	Use of multi-directional projected depth motion maps (MPDMM).	MSRAction3D [107] MSRGesture3D [108]

2.3. Deep Learning Based Methods

After being a breakthrough in image classification, it was a matter of time to start using deep learning for video-based activity recognition. Although great advances have been made and state-of-the-art results have been achieved, the level of image classification has not been reached yet.

In 2014, a paper was released [110] encouraged by the results of *Convolutional Neural Networks* (*CNNs*) [111] for image recognition problems [112–115]. Using a 1M videos dataset, they studied different ways for extending the connectivity of a CNN in a time domain in order to take advantage of local spatio-temporal information. They proposed three connectivity patterns: Early Fusion, Late Fusion and Slow Fusion. The Early Fusion extension combines information across an entire time window immediately on the pixel level. The Late Fusion model places two separate single-frame networks with shared parameters a distance of 15 frames apart and then merges the two streams in the first fully connected layer. This way motion can not be detected until the fully connected layer, which compares both outputs to compute global motion. The Slow Fusion model slowly fuses temporal information throughout the network such that higher layers get access to progressively more global information in both spatial and temporal dimensions. For optimization, Downpour Stochastic Gradient Descent [116] is used. The results show that a slow fusion model performs better than the early and late fusion alternatives. They also find out that a single-frame model already displays very strong performance, suggesting that local motion may not be critically important.

In the same year as the previous paper, another work was published [117] that has been the reference of later publications. Simonyan et al. propose a two-stream Convolutional Neural Network architecture that incorporates spatial and temporal networks. Videos can naturally be decomposed into spatial and temporal components. The spatial part provides information about scenes and objects of the video, taking as input a single frame. Nevertheless, the temporal part, which consists of stacked optical flow vectors, shows the movement of the observer (the camera) and the objects in the form of motion across the frames. This way, the authors divide the architecture into two streams. Each stream is implemented using a deep ConvNet [118]; softmax scores are combined by late fusion using a SVM [119] or averaging. It seems that training a temporal network with optical flow improves the training of just stacked frames as in [110]. However, compared to the shallow representation of [76], there are some things to improve yet.

After these two publications, and taking them as a starting point, deep learning has continued to be used for activity recognition, mainly with Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) [120].

In [121], the authors investigate if recurrent models are effective for tasks involving sequences. They propose a Long-term Recurrent Convolutional Network (LRCN) and demonstrate the value of these models for activity recognition. The LSTM unit they use is as the one described in [122]. Compared to previous models, recurrent convolutional models learn compositional representations in space and time and not just assume a fixed visual representation or perform simple temporal averaging for sequential processing. As input, both RGB and optical flow are used and it is observed that the best results are achieved by the weighted scores of both inputs as in [117]. They show that learning sequential dynamics with a deep sequence model improves previous methods that only took into account parameters of the visual domain.

Wang et al. in their work [123] presented very deep two-stream ConvNets in order to improve the results of recent architectures [117] getting closer to image domain deep models. Apart from using two known architectures, GoogLeNet [124] and VGGNet-16 [125], they use 10-frame stacking of optical flow for the temporal network and a single frame image for the spatial network. As the training datasets are small, the model is initialized by pre-training it with ImageNet and, to avoid over-fitting, dropout and data augmentation techniques are used. They proposed two new data augmentation methods: one of them consists of cropping four corners and one center of the images and, in the other, multi-scale cropping is used.

In [126], trajectory-pooled deep-convolutional descriptor (TDD) is introduced, which combines the works of [76,117]. The authors first train two-stream ConvNets and use them as feature extractors to achieve convolutional spatial and temporal feature maps from the learned networks. With the improved trajectories method, a set of point trajectories are detected and, using trajectory pooling, TDD descriptors are created based on normalized convolutional feature maps and these trajectories, as in Equation (5):

$$D(T_k, \widetilde{C}^a_m) = \sum_{p=1}^{P} \widetilde{C}^a_m(\overline{(r_m \times x^k_p)}, \overline{(r_m \times y^k_p)}, z^k_p), \tag{5}$$

where T_k is a trajectory, $\widetilde{C}^a_m$ is a mth layer normalized feature map, (x^k_p, y^k_p, z^k_p) is the pth point position of video coordinates of T_k trajectory and r_m is the mth layer map size ratio, $\overline{(\cdot)}$ being the rounding operation. Fisher vector representation is used to bring together TDDs over the whole video and, finally, an SVM classifier does the recognition.

Although having some similarities with previous works [110,117], in [127], Tran et al., instead of using 2D convolutions across frames, use 3D convolutions and 3D pooling, propagating temporal information across all the layers in the network. They propose a simple yet effective approach for spatio-temporal feature learning using deep three-dimensional, convolutional networks trained on a large scale supervised video dataset. They show that 3D ConvNets [110,128] with a linear classifier are more suitable for spatio-temporal feature learning than 2D ConvNets and that the model performs even better additionally using hand-crafted features like iDT [76].

In the work by Feichtenhofer et al. [129], the authors add two ideas to the two-stream architecture of [117]. They show that it is important to associate spatial feature maps of a particular area to temporal feature maps for that corresponding region. The spatial and temporal networks are fused at an early level, so, rather than fusing at the softmax layer, they are fused at a convolutional layer. The fusion can be made in different ways and, in [130], Yue-Hei et al. evaluate many other methods to combine two-stream ConvNets over time. The architecture they propose does not increase the number of parameters significantly compared to previous methods and their results are improved by adding also iDT features [76].

Wang et al. also improved the two streams architecture in their work [131], presenting a long-rate temporal structure model, the Temporal Segment Network (TSN). Most of the previous works were not able to incorporate long-range temporal structures, but their model combines a sparse temporal sampling strategy and video-level supervision to enable efficient and effective learning using the whole action video. Another problem they wanted to deal with was over-fitting because, due to the difficulty of collecting data, the available datasets were limited. They use different techniques to avoid the risk of over-fitting: batch normalization [132], dropout [133] and pre-training. The authors also evaluate the model using four different input modalities: optical-flow, warped optical-flow, RGB and RGB difference, the last one inspired by [134].

Bilen et al. [135] presented the concept of dynamic image, which summarizes a video into just a single RGB image by applying rank pooling on the images of a video. This way, image classification CNNs can be used directly, as the input is an image. The idea of reducing the whole video to a single image is taken from [136]. In their experiments, two scenarios were considered: getting a single dynamic figure from a video or getting several dynamic images from each video, the second approach is thought to deal with the lack of training videos. Then, dynamic feature maps are obtained by adding a new temporal layer to the CNN and a pre-trained CaffeNet [137] model is used to initialize the network.

In 2017, Carreira et al. [138] presented a new architecture that uses two different 3D networks for both streams of a two-stream architecture [117], called Two-Stream Inflated 3D ConvNet (I3D). It is based on 2D ConvNet inflation, expanding filters and pooling kernels of very deep image classification ConvNets into 3D, leading to very deep spatio-temporal classifiers and making it possible to learn spatio-temporal feature extractors from videos. In basic two-stream architectures the spatial stream is formed by single frames; however, in I3D, the spatial stream input consists of frames stacked in time dimension. Apart from the new model, the main contribution of this paper is a new dataset for action recognition, the Kinetics Human Action Video dataset, which is two orders of magnitude larger than previous datasets with 400 actions and more than 400 clips per action collected from YouTube. They also showed that, when pre-training on Kinetics, results of I3D models are improved.

Later in 2018, [139] improved the performance of [121] by using lower spatial resolution and longer clips to keep the complexity of networks tractable while dealing with the inability to capture long range temporal information. They consider space-time convolutional neural networks [127,128,140] and study architectures with long-term temporal convolutions (LTC), which are used to learn video representations. As in [121], different low-level representations are studied: RGB and optical flow. Their experiments confirm the advantage of motion-based representations and highlight the importance of good quality motion estimation for learning efficient representations for human action recognition.

Ullah et al. [141] proposed an action recognition method by processing the video data using convolutional neural networks (CNN) and deep bidirectional LSTM (DB-LSTM) networks [142]. On the one hand, in order to reduce complexity and redundancy, deep features are extracted from every six frames of a video using pre-trained AlexNet [112]. Then, the sequential information among frame features is learned using an DB-LSTM network, where multiple layers are stacked together in both forward pass and backward pass of DB-LSTM to increase its depth. The video is analyzed in N chunks and N depends on processing time interval T. The final output is the combination of small chunks outputs. As the video is processed and features are analyzed for a certain time interval, the proposed method is able to learn long sequences and recognize actions in long videos.

Wang et al. [143] proposed a discriminative pooling based on the idea that, among the frames, not all of them have the same importance and a few are those that provide characteristic information about the action [144]; some of the features in one sequence are indeed useful, while the rest are not. Taking all the CNN features as positive (containing good and bad features) and the known background or noisy frames as negative, a nonlinear hyperplane that differentiates the discriminative features from the rest is learned to make the separation. The decision boundary of the classifier thus learned is then used as a descriptor for the entire video sequence, which they call the SVM Pooled (SVMP) descriptor. Thus, they formulate an efficient solver that learns these hyperplanes per video and the corresponding action classifiers over the hyperplanes. This pooling scheme is end-to-end trainable within a deep framework.

The authors of [145] presented the first end-to-end convNets which admit videos of arbitrary size and length. After seeing that 3D convolutional networks have achieved good results in action recognition, they decided to delete two of the requirements that existing convNets had: fixed size and length input videos were required, which reduce the quality of video analysis. Basically, each video is decomposed into spatial and temporal shots and, for both pieces of information, the same process is computed. A spatial temporal pyramid pooling (STPP) convNet is first used to extract equal-dimensional descriptors from variable-sized frame sequences. Then, a Long Short-Term Memory (LSTM) or a CNN-E model is used to recognize the actions from these descriptors. Finally, both streams (spatial and temporal) are combined by a late fusion.

In Table 3, a summary of the deep learning based methods explained is presented.

Table 3. Summary of deep learning based methods.

	YEAR	SUMMARY	DATASET
Karpathy et al. [110]	2014	Use of different connectivity patterns for CNNs: early fusion, late fusion and slow fusion.	Sports-1M [110] UCF101 [89]
Simonyan et al. [117]	2014	Use of a two-stream CNN architecture, incorporating spatial and temporal networks.	UCF101 [89] HMDB51 [88]
Donahue et al. [121]	2015	Use of a Long-term Recurrent Convolutional Network (LRCN) to learn compositional representations in space and time.	UCF101 [89]
Wang et al. [123]	2015	Use of very deep two-stream convNets, using stacked optical flow for temporal network and a single frame image for spatial network.	UCF101 [89]
Wang et al. [126]	2015	Use of trajectory-pooled deep-convolutional descriptor (TDD).	UCF101 [89] HMDB51 [88]
Tran et al. [127]	2015	Use of deep 3D convolutional networks, which are better for spatio-temporal feature learning.	UCF101 [89]
Feichtenhofer et al. [129]	2016	Use of two-stream architecture associating spatial feature maps of a particular area to temporal feature maps of that region and fusing the networks at an early level.	UCF101 [89] HMDB51 [88]
Wang et al. [131]	2016	Use of Temporal Segment Network (TSN) to incorporate long-range temporal structures avoiding overfitting.	UCF101 [89] HMDB51 [88]
Bilen et al. [135]	2016	Use of image classification CNNs after summarizing the videos in dynamic images.	UCF101 [89] HMDB51 [88]
Carreira et al. [138]	2017	Use of two-stream Inflated 3D ConvNet (I3D), using two different 3D networks for both streams of a two-stream architecture.	UCF101 [89] HMDB51 [88]
Varol et al. [139]	2018	Use of space-time CNNs and architectures with long-term temporal convolutions (LTC), using lower spatial resolution and longer clips.	UCF101 [89] HMDB51 [88]
Ullah et al. [141]	2018	Use of CNNs to reduce complexity and redundancy and deep bidirectional LSTM (DB-LSTM) to learn sequential information among frame features.	UCF101 [89] HMDB51 [88] YouTube actions [146]
Wang et al. [143]	2018	Use of a discriminative pooling, taking into account that just a few frames provide characteristic information about the action.	HMDB51 [88]
Wang et al. [145]	2018	Use of convNets which admit videos of arbitrary size and length, using first a STPP and a LSTM (or CNN-E) then.	UCF101 [89] HMDB51 [88] ACT [147]

Finally, in order to compare the presented techniques briefly, some advantages and disadvantages are presented in Table 4.

Table 4. Advantages and disadvantages of presented techniques.

	Advantages	Disadvantages
Hand-crafted motion features	- There is no need of a large amount of data for training. - It is simple and unambiguous to understand the model and analyze and visualize the functions. - The features used to train the model are explicitly known.	- Usually these features are not robust. - They can be computationally intensive due to the high dimensions. - The discriminative power is usually low.
Depth information	- The 3D structure information of the image that depth sensors provide is used to recover postures and recognize the activity. - The skeletons extracted from depth maps are precise. - Depth sensors can work in darkness.	- Depth maps have no texture, making it difficult to apply local differential operators. - The global features can be unsettled because depth maps may contain occlusions.
Deep Learning	- There is no need of expert knowledge to get suitable features, reducing the effort of feature extraction. - Instead of designing them manually, features are automatically learned through the network. - Deep neural networks can extract high-level representation in deep layer, making it more suitable for complex tasks.	- Need to collect massive data, consequently there is a lack of data sets. - Time consuming. - Problem of models capability of generalization.

3. Benchmark Datasets

Although there is not a standard benchmark in activity recognition, there are some datasets that are being considered as reference ones [148]. As it has been mentioned above, due to the complexity of collecting data, the available datasets are limited. In this section, the most used datasets are presented.

3.1. UCF-101

UCF101 [89,149] is an action recognition dataset of realistic action videos. It is composed of 13,320 videos with 101 action categories and 27 h of video data. This dataset is an extension of the UCF50 [150] dataset that has 50 action categories.

The videos have been collected from YouTube, making the dataset realistic, and it provides a great variety of videos with different objects, camera motion, background, lighting, viewpoint, etc. Based on those features, videos are gathered into 25 groups (4–7 videos per action in each group) with videos sharing some of the features, as background, for example.

The 101 categories can be divided in five main groups:

1. Human–Object Interaction: twenty categories.
2. Body-Motion Only: sixteen categories.
3. Human–Human Interaction: five categories.
4. Playing Musical Instruments: ten categories.
5. Sports: fifty categories.

3.2. HMDB51

HMDB51 [88,151] is another action recognition database that collects videos from various sources, mainly from movies but also from public databases such as YouTube, Google and Prelinger Archives.

It consists of 6849 videos with 51 action categories and a minimum of 101 clips belong to each category. The action categories can be divided as well in five main groups:

1. General facial actions: smile, laugh, chew, talk.
2. Facial actions with object manipulation: smoke, eat, drink.
3. General body movements: cartwheel, clap hands, climb, climb stairs, dive, fall on the floor, backhand flip, handstand, jump, pull up, push up, run, sit down, sit up, somersault, stand up, turn, walk, wave.
4. Body movements with object interaction: brush hair, catch, draw sword, dribble, golf, hit something, kick ball, pick, pour, push something, ride bike, ride horse, shoot ball, shoot bow, shoot gun, swing baseball bat, sword exercise, throw.
5. Body movements for human interaction: fencing, hug, kick someone, kiss, punch, shake hands, sword fight.

Apart from the action label, other meta-labels are indicated in each clip. These labels provide information about some features describing properties of the clip, such as camera motion, lighting conditions, or background. As videos are taken from movies or YouTube, the variation of features is high and that extra information can be useful. In addition, the quality of the videos has been measured (*good*, *medium*, *bad*), and they are rated depending on whether body parts vanish while action is executed or not.

3.3. Weizmann

Before the two previous databases were created, many methods used the Weizmann [152] database published by [64] to evaluate the performance of their contributions. It provides 90 low-resolution (180 × 144, deinterlaced 50 fps) video sequences. These clips show 10 different actions performed by nine different people. These are the actions that appear in the database: *run, walk, skip, jumping-jack (jack), jump-forward-on-two-kegs (jump), jump-in-place-on-two-legs (pjump), side-gallop (side), wave-two-hands (wave2), wave-one-hand (wave1)* and *bend*. Background and the viewpoint are statics.

3.4. MSRAction3D

In 2010, as there was no public benchmark database, the authors in [107] published the database called MSRAction3D [153] which provided the sequences of depth maps captured by a depth camera. The dataset contains twenty actions: *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up* and *throw*. Seven different individuals performed each action three times, facing the camera during the performance. The depth maps have a size of 640 × 480 and they were captured at about 15 frames per second (fps) by a depth camera with infra-red light structure.

3.5. ActivityNet

The authors of [154] presented in 2015 the ActivityNet [155] database. It is composed of 203 different classes with an average of 137 videos per class and a total of 648 video hours. The videos were obtained from online video sharing sites and they are around 5–10 min long. Half of the videos are in HD resolution (1280 × 720) and most of them have a frame rate of 30 fps.

The aim of this database is to collect activities of humans daily life and it has a hierarchical structure, organizing the activities according to social interactions and where they take place.

3.6. Something Something

Later, in 2017, the authors of [156] introduced the "Something Something" [157] dataset. The first version of the database consists of 108,499 videos belonging to 174 different labels with 23,137 distinct object names. The length of the videos variate between 2 and 6 s and they have a height of 100 px and variable width. Labels are textual descriptions such as "Putting *something* next to *something*" where *something* refers to an object name. This database is already split into train, validation and test, containing 86,017, 11,522 and 10,960 videos, respectively.

However, there has been a second release of the dataset and now it contains 220,847 videos, 168,913 for the training set, 24,777 for the validation set and 27,157 for the test set. The number of labels remains the same, but there are additional object annotations now. Moreover, the pixel resolution has increased from 100 px to 240 px.

3.7. Sports-1M

In [110], Karpathy et al. presented a new database, Sports-1M [158], which contains 1,133,158 video URLs with 487 automatically annotated different labels. YouTube Topics API was used to do the annotation. There are around 1000–3000 videos per class and some of them, nearly the 5%, are labelled with more than one class.

Nowadays, the YouTube-8M [159] dataset is also available and the Sports-1M dataset is included in it. This dataset is composed of videos from 3862 labels and it contains 350,000 h of video. In this case, each video has an average of three labels.

3.8. AVA

The authors of [160] presented AVA [161], a video dataset of spatio-temporally localized Atomic Visual Actions. This dataset consists of 430 movie clips of 15 min length annotated with 80 actions (14 poses, 17 person–person, 49 person–object). There are 386,000 labelled segments, 614,000 labelled bounding boxes and 81,000 person tracks, with a total of 1.58M labelled actions, with multiple labels per person occurring frequently.

Every person of the scene is localized by a bounding box and labels are assigned according to the action performed by the actor. Each scene can have more than a label, one of them corresponds to the actor's pose and additional labels which correspond to person–object or person–person interactions can be assigned. A frame containing more than one actor is labelled separately for each person of the scene.

To finish, in Table 5, a summary of the explained datasets is introduced, in order to present the information more clearly.

Table 5. Summary of the presented datasets.

	# Classes	**# Videos**	**# Actors**	**Resolution**	**Year**
Weizmann	10	90	9	180 × 144	2005
MSRAction3D	20	420	7	640 × 480	2010
HMDB51	51	6849	-	320 × 240	2011
UCF50	50	6676	-	-	2012
UCF101	101	13,320	-	320 × 240	2012
Sports-1M	487	1,133,158	-	-	2014
ActivityNet	203	27,801	-	1280 × 720	2015
Something Something	174	220,847	-	__ (Variable width) × 240	2017
AVA	80	430	-	-	2018

4. Results

To better analyze the explained methods and the contributions of each one of them, the results obtained for mentioned datasets are compared. For each method, the achieved accuracy values for different datasets are shown, together with the reference to the original article where they have been proposed.

On the one hand, in Table 6, results for depth information based methods can be observed. These methods use the MSRAction3D as benchmarks because the input they need is different from other models. Regarding the methods used with the MSRAction3D database, the best result of the presented methods is achieved by [102], as it can be seen in Table 6.

Table 6. Obtained accuracies for the benchmark dataset with depth information based methods.

	METHOD	**MSRAction3D**
DS	DMM-HOG [98]	85.52%
	HON4D [99]	88.89%
	M3DLSK+STV [102]	**95.36%**
	MPDMM [104]	94.8%

On the other hand, most of the hand-crafted feature methods use the Weizmann dataset as a benchmark. However, some of the presented models work with both UCF101 and HMDB51 datasets, which are used as benchmarks in deep learning methods. Thus, in Table 7, the obtained accuracy values can be observed, together for deep learning and hand-crafted methods.

Table 7. Obtained accuracies for the benchmark datasets with hand-crafted methods and deep learning methods.

	METHOD	**UCF101**	**HMDB51**	**Weizmann**
Hand-crafted	Hierarchical [55]	-	-	72.8%
	Far Field of View [63]	-	-	**100%**
	HOOF NLDS [46]	-	-	94.4%
	Direction HOF [45]	-	-	79.17%
	iDT [76]	-	57.2%	-
	iDT+FV [76]	85.9%	57.2%	-
	OF Based [81]	-	-	90.32%
	Edges OF [15]	-	-	95.69%
	HOG features [87]	-	-	99.7%
Deep learning	Slow Fusion CNN [110]	65.4%	-	-
	Two stream (avg) [117]	86.9%	58.0%	-
	Two stream (SVM) [117]	88.0%	59.4%	-
	IDT+MIFS [162]	89.1%	65.1%	-
	LRCN (RGB) [121]	68.2%	-	-
	LRCN (FLOW) [121]	77.28%	-	-
	LRCN (avg, 1/2-1/2) [121]	80.9%	-	-
	LRCN (avg, 1/3-2/3) [121]	82.34%	-	-
	Very deep two-stream (VGGNet-16) [123]	91.4%	-	-
	TDD [126]	90.3%	63.2%	-
	TDD + iDT [126]	91.5%	65.9%	-
	C3D [127]	85.2%	-	-
	C3D + iDT [127]	90.4%	-	-
	TwoStreamFusion [129]	92.5%	65.4%	-
	TwoStreamFusion+iDT [129]	93.5%	69.2%	-
	TSN (RGB+FLOW) [131]	94.0%	68.5%	-
	TSN (RGB+FLOW+WF) [131]	94.2%	69.4%	-
	Dynamic images + iDT [135]	89.1%	65.2%	-
	Two-StreamI3D [138]	93.4%	66.4%	-
	Two-StreamI3D, pre-trained [138]	**97.9%**	80.2%	-
	LTC (RGB) [139]	82.4%	-	-
	LTC (FLOW) [139]	85.2%	59.0%	-
	LTC(FLOW+RGB) [139]	91.7%	64.8%	-
	LTC(FLOW+RGB)+iDT [139]	92.7%	67.2%	-
	DB-LSTM [141]	91.21%	**87.64%**	-
	Two-Stream SVMP(VGGNet) [143]	-	66.1%	-
	Two-Stream SVMP(ResNet) [143]	-	71.0%	-
	Two-Stream SVMP(+ iDT) [143]	-	72.6%	-
	Two-Stream SVMP(I3D conf) [143]	-	83.1%	-
	STPP + CNN-E (RGB) [145]	85.6%	62.1%	-
	STPP + LSTM (RGB) [145]	85.0%	62.5%	-
	STPP + CNN-E (FLOW) [145]	83.2%	55.4%	-
	STPP + LSTM (FLOW) [145]	83.8%	54.7%	-
	STPP + CNN-E (RGB+FLOW) [145]	92.4%	70.5%	-
	STPP + LSTM (RGB+FLOW) [145]	92.6%	70.3%	-

As it can be seen in Table 7, the Two-Stream I3D method [138], pre-trained with the Kinetics dataset, provides the best result for the UCF101 dataset. For HMDB51, the best result is achieved by the DB-LSTM model [141] and, among those who have tested with the Weizmann database, the best value is given by the method presented in [63].

5. Discussion

After reading the previous sections, the researchers could ask themselves which are the most promising lines of research in the field of action recognition in videos, or where it is more likely to get a higher return for the invested effort.

For people just interested in applying an existing method to their data, or in minimal modifications or customizations, some authors of the presented methods have made their code available that can be used. These implementations are indicated in Table 8. Among them, the methods explained in [138,141] provide the best results.

Table 8. Available code for presented methods.

METHOD	YEAR	PAPER	CODE
Deep Learning	2018	Video representation learning using discriminative pooling [143]	SVMP https://github.com/3xWangDot/SVMP
Deep Learning	2018	Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features [141]	Bi-directional LSTM https://github.com/Aminullah6264/BidirectionalLSTM
Deep Learning	2018	Long-term temporal convolutions for action recognition [139]	LTC https://github.com/gulvarol/ltc
Deep Learning	2017	Quo vadis, action recognition? A new model and the Kinetics dataset [138]	Two-Stream I3D https://github.com/deepmind/kinetics-i3d
Deep Learning	2016	Dynamic image networks for action recognition [135]	Dynamic images https://github.com/hbilen/dynamic-image-nets
Deep Learning	2016	Temporal segment networks: Towards good practices for deep action recognition [131]	TSN https://github.com/yjxiong/temporal-segment-networks
Deep Learning	2016	Convolutional two-stream network fusion for video action recognition [129]	Two-Stream Fusion https://github.com/feichtenhofer/twostreamfusion
Deep Learning	2015	Learning spatiotemporal features with 3D convolutional networks [127]	C3D https://github.com/facebook/C3D
Deep Learning	2015	Action recognition with trajectory-pooled deep-convolutional descriptors [126]	TDD https://github.com/wanglimin/tdd/
Deep Learning	2015	Towards good practices for very deep two-stream convNets [123]	Very deep Two-Stream convNets https://github.com/yjxiong/caffe/tree/action_recog
Depth information	2013	HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences [99]	HON4D http://www.cs.ucf.edu/~oreifej/HON4D.html
Hand-crafted motion features	2013	Action Recognition with Improved Trajectories [76]	Improved Trajectories http://lear.inrialpes.fr/~wang/improved_trajectories

For researchers interested in developing new methods or in deep modifications of current ones, deep learning looks like the way to go, although the computations costs could be forbidding.

Some researchers could have the resources to generate new datasets similar to those presented in this paper. They need to have in mind several decisions: will the videos have the same resolution and/or length and will they be recorded with the same background? We advocate for datasets covering different kinds of videos, with that information present in the metadata and with enough samples of each type. Anyway, if the researcher resources are somewhat limited, which is usually the case, it is advisable to focus on just one type of video. All the technical information about the sensor might appear in the metadata, as well as lighting conditions or any information of interest. If depth sensors are used, high and low resolution of the depth data could be provided. Processing of depth data can be computationally expensive, and other researchers using that dataset could benefit from access to a standard low resolution version of that data.

The task of labeling the database samples can be eased with the help of some tools. While just providing a global label for a video does not require a great deal of effort, the video database curators could choose to gather information about individual frames in the videos. There are several tools that could be useful in this task, like Sloth [163], LabelMe [164,165] or LabelBox [166].

It is difficult to predict the future development of this area, but, at least in the short term, the overall tendency in machine learning is going towards massive data, computationally expensive algorithms and dedicated hardware. It is expected that the price of depth sensors will keep a descending curve, as well as the cost of hardware in general. The main challenges are expected to be twofold: for the researchers developing new methods, those related to the storage and processing of massive databases, and for developers integrating the methods into software solutions, those related to a fast classification time.

6. Conclusions

In this paper, different methods for video activity recognition have been presented. Several models have been explained showing the development of recent years. Likewise, several databases used to evaluate the performance of the models have been introduced. The results have been shown together in a table in order to compare the methods presented correctly.

Due to the extent width of the subject, there are many more models that have not been mentioned in this document. Even so, an attempt has been made to show a current state-of-the-art by presenting different techniques to deal with the problem. To sum up, through this document, we have tried to show the relevance and current situation of video-based activity recognition.

Video-based activity recognition, as it has been mentioned before, is more complicated than static image classification and this is also reflected in the results obtained so far. However, since deep learning is still being exploited, in the near future, this task may become easier to perform and current results may be improved using some deep learning techniques.

Author Contributions: B.S. and J.M.M.-O. supervised the paper. I.R.-M. collected the materials and wrote the paper. I.R. and E.J. revised and formatted the paper.

Funding: This work has been partially supported by the Basque Government, Spain (IT900-16), the Spanish Ministry of Economy and Competitiveness (RTI2018-093337-B-I00 , MINECO/FEDER, EU).

Acknowledgments: The authors thank the editor and three anonymous reviewers for their comments on the early version of this manuscript.

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

References

1. Avci, A.; Bosch, S.; Marin-Perianu, M.; Marin-Perianu, R.; Havinga, P. Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey. In Proceedings of the 23th International Conference on Architecture of Computing Systems 2010, Hannover, Germany, 22–23 February 2010; pp. 1–10.
2. Mulroy, S.; Gronley, J.; Weiss, W.; Newsam, C.; Perry, J. Use of cluster analysis for gait pattern classification of patients in the early and late recovery phases following stroke. *Gait Posture* **2003**, *18*, 114–125. [CrossRef]
3. Rautaray, S.S.; Agrawal, A. Vision based hand gesture recognition for human computer interaction: A survey. *Artif. Intell. Rev.* **2015**, *43*, 1–54. [CrossRef]
4. Mitra, S.; Acharya, T. Gesture recognition: A survey. *IEEE Trans. Syst. Man Cybern. Part Appl. Rev.* **2007**, *37*, 311–324. [CrossRef]
5. Vishwakarma, S.; Agrawal, A. A survey on activity recognition and behavior understanding in video surveillance. *Vis. Comput.* **2013**, *29*, 983–1009. [CrossRef]
6. Leo, M.; D'Orazio, T.; Spagnolo, P. Human activity recognition for automatic visual surveillance of wide areas. In Proceedings of the ACM 2nd International Workshop on Video Surveillance & Sensor Networks, New York, NY, USA, 15 October 2004; pp. 124–130.
7. Coppola, C.; Cosar, S.; Faria, D.R.; Bellotto, N. Social Activity Recognition on Continuous RGB-D Video Sequences. *Int. J. Soc. Robot.* **2019**, 1–15. [CrossRef]
8. Coppola, C.; Faria, D.R.; Nunes, U.; Bellotto, N. Social activity recognition based on probabilistic merging of skeleton features with proximity priors from RGB-D data. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea, 9–14 October 2016; pp. 5055–5061.
9. Lin, W.; Sun, M.T.; Poovandran, R.; Zhang, Z. Human activity recognition for video surveillance. In Proceedings of the 2008 IEEE International Symposium on Circuits and Systems, Seattle, WA, USA, 18–21 May 2008; pp. 2737–2740.
10. Nair, V.; Clark, J.J. Automated visual surveillance using Hidden Markov Models. *International Conference on Vision Interface*. 2002, pp. 88–93. Available online: https://pdfs.semanticscholar.org/8fcf/7e455419fac79d65c62a3e7f39a945fa5be0.pdf (accessed on 15 July 2019).
11. Ma, M.; Meyer, B.J.; Lin, L.; Proffitt, R.; Skubic, M. VicoVR-Based Wireless Daily Activity Recognition and Assessment System for Stroke Rehabilitation. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 3–6 December 2018; pp. 1117–1121.
12. Ke, S.R.; Thuc, H.; 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]
13. 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]
14. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **2017**, *60*, 4–21. [CrossRef]
15. Kumar, S.S.; John, M. Human activity recognition using optical flow based feature set. In Proceedings of the 2016 IEEE International Carnahan Conference on Security Technology (ICCST), Orlando, FL, USA, 24–27 October 2016; pp. 1–5.
16. Guo, K.; Ishwar, P.; Konrad, J. Action recognition using sparse representation on covariance manifolds of optical flow. In Proceedings of the 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, Boston, MA, USA, 29 August–1 September 2010; pp. 188–195.
17. Niu, F.; Abdel-Mottaleb, M. HMM-based segmentation and recognition of human activities from video sequences. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005; pp. 804–807.
18. Raman, N.; Maybank, S.J. Activity recognition using a supervised non-parametric hierarchical HMM. *Neurocomputing* **2016**, *199*, 163–177. [CrossRef]
19. Liciotti, D.; Duckett, T.; Bellotto, N.; Frontoni, E.; Zingaretti, P. HMM-based activity recognition with a ceiling RGB-D camera. In Proceedings of the ICPRAM—6th International Conference on Pattern Recognition Applications and Methods, Porto, Portugal, 24–26 February 2017.
20. Ma, M.; Fan, H.; Kitani, K.M. Going deeper into first-person activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1894–1903.

21. Nunez, J.C.; Cabido, R.; Pantrigo, J.J.; Montemayor, A.S.; Velez, J.F. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognit.* **2018**, *76*, 80–94. [CrossRef]
22. 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, Providence, RI, USA, 16–21 June 2012; pp. 1234–1241.
23. Ng, J.Y.H.; Davis, L.S. Temporal difference networks for video action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*; IEEE: Piscataway, NJ, USA, 2018; pp. 1587–1596.
24. Lan, T.; Sigal, L.; Mori, G. Social roles in hierarchical models for human activity recognition. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1354–1361.
25. Vahora, S.; Chauhan, N. Deep neural network model for group activity recognition using contextual relationship. *Eng. Sci. Technol. Int. J.* **2019**, *22*, 47–54. [CrossRef]
26. Huang, S.C. An advanced motion detection algorithm with video quality analysis for video surveillance systems. *IEEE Trans. Circuits Syst. Video Technol.* **2010**, *21*, 1–14. [CrossRef]
27. Hu, W.; Tan, T.; Wang, L.; Maybank, S. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Syst. Man Cybern. Part Appl. Rev.* **2004**, *34*, 334–352. [CrossRef]
28. Gaba, N.; Barak, N.; Aggarwal, S. Motion detection, tracking and classification for automated Video Surveillance. In Proceedings of the 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), Delhi, India, 4–6 July 2016; pp. 1–5.
29. Trucco, E.; Plakas, K. Video tracking: A concise survey. *IEEE J. Ocean. Eng.* **2006**, *31*, 520–529. [CrossRef]
30. Maggio, E.; Cavallaro, A. *Video Tracking: Theory and Practice*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
31. Del Rincón, J.M.; Santofimia, M.J.; Nebel, J.C. Common-sense reasoning for human action recognition. *Pattern Recognit. Lett.* **2013**, *34*, 1849–1860. [CrossRef]
32. Santofimia, M.J.; Martinez-del Rincon, J.; Nebel, J.C. Episodic reasoning for vision-based human action recognition. *Sci. World J.* **2014**, *2014*. [CrossRef]
33. Onofri, L.; Soda, P.; Pechenizkiy, M.; Iannello, G. A survey on using domain and contextual knowledge for human activity recognition in video streams. *Expert Syst. Appl.* **2016**, *63*, 97–111. [CrossRef]
34. Wang, X.; Gao, L.; Song, J.; Zhen, X.; Sebe, N.; Shen, H.T. Deep appearance and motion learning for egocentric activity recognition. *Neurocomputing* **2018**, *275*, 438–447. [CrossRef]
35. Aggarwal, J.K.; Ryoo, M.S. Human activity analysis: A review. *ACM Comput. Surv. (CSUR)* **2011**, *43*, 16. [CrossRef]
36. Kong, Y.; Fu, Y. Human Action Recognition and Prediction: A Survey. *arXiv* **2018**, arXiv:1806.11230.
37. Raptis, M.; Sigal, L. Poselet key-framing: A model for human activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2650–2657.
38. Wang, Y.; Sun, S.; Ding, X. A self-adaptive weighted affinity propagation clustering for key frames extraction on human action recognition. *J. Vis. Commun. Image Represent.* **2015**, *33*, 193–202. [CrossRef]
39. Niebles, J.C.; Wang, H.; Fei-Fei, L. Unsupervised learning of human action categories using spatial-temporal words. *Int. J. Comput. Vis.* **2008**, *79*, 299–318. [CrossRef]
40. Dollár, P.; Rabaud, V.; Cottrell, G.; Belongie, S. 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; pp. 65–72.
41. Bregonzio, M.; Gong, S.; Xiang, T. Recognising action as clouds of space-time interest points. In Proceedings of the CVPR 2009, Miami Beach, FL, USA, 20–25 June 2009; Volume 9, pp. 1948–1955.
42. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008, pp. 1–8.
43. Ngo, C.W.; Pong, T.C.; Zhang, H.J. Motion-based video representation for scene change detection. *Int. J. Comput. Vis.* **2002**, *50*, 127–142. [CrossRef]
44. Sand, P.; Teller, S. Particle video: Long-range motion estimation using point trajectories. *Int. J. Comput. Vis.* **2008**, *80*, 72. [CrossRef]

45. Lertniphonphan, K.; Aramvith, S.; Chalidabhongse, T.H. Human action recognition using direction histograms of optical flow. In Proceedings of the 2011 11th International Symposium on Communications & Information Technologies (ISCIT), Hangzhou, China, 12–14 October 2011; pp. 574–579.
46. Chaudhry, R.; Ravichandran, A.; Hager, G.; Vidal, R. Histograms of oriented optical flow and Binet–Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1932–1939.
47. 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]
48. Bobick, A.; Davis, J. An appearance-based representation of action. In Proceedings of the 1996 International Conference on Pattern Recognition (ICPR '96), Washington, DC, USA, 25–30 August, 1996; pp. 307–312.
49. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04), Washington, DC, USA, 23–26 August 2004; pp. 32–36.
50. Laptev, I. On space-time interest points. *Int. J. Comput. Vis.* **2005**, *64*, 107–123. [CrossRef]
51. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.
52. Vapnik, V.N. An overview of statistical learning theory. *IEEE Trans. Neural Netw.* **1999**, *10*, 988–999. [CrossRef] [PubMed]
53. Wallraven, C.; Caputo, B.; Graf, A. Recognition with local features: The kernel recipe. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 3–16 October 2003; p. 257.
54. Wof, L.; Shashua, A. Kernel principal angles for classification machines with applications to image sequence interpretation. In Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 8–20 June 2003.
55. Niebles, J.C.; Fei-Fei, L. A hierarchical model of shape and appearance for human action classification. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
56. Bouchard, G.; Triggs, B. Hierarchical part-based visual object categorization. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; pp. 710–715.
57. Bosch, A.; Zisserman, A.; Munoz, X. Representing shape with a spatial pyramid kernel. In Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands, 9–11 July 2007; pp. 401–408.
58. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 7–22 June 2006; pp. 2169–2178.
59. Marszałek, M.; Schmid, C.; Harzallah, H.; Van De Weijer, J. Learning object representations for visual object class recognition. In Proceedings of the Visual Recognition Challange Workshop, in Conjunction with ICCV, Rio de Janeiro, Brazil, October 2007. Available online: https://hal.inria.fr/inria-00548669/ (accessed on 15 July 2019).
60. Zhang, J.; Marszałek, M.; Lazebnik, S.; Schmid, C. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int. J. Comput. Vis.* **2007**, *73*, 213–238. [CrossRef]
61. Harris, C.; Stephens, M. A combined corner and edge detector. In Proceedings of the 4th Alvey Vision Conference, Manchester, UK, 31 August–2 September 1988; pp. 10–5244.
62. Horn, B.K.; Schunck, B.G. Determining optical flow. *Artif. Intell.* **1981**, *17*, 185–203. [CrossRef]
63. Chen, C.C.; Aggarwal, J. Recognizing human action from a far field of view. In Proceedings of the 2009 Workshop on Motion and Video Computing (WMVC), Snowbird, UT, USA, 8–9 December 2009; pp. 1–7.
64. Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time shapes. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Beijing, China, 17–21 October 2005; pp. 1395–1402.
65. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 886–893.

66. Hatun, K.; Duygulu, P. Pose sentences: a new representation for action recognition using sequence of pose words. In Proceedings of the 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–4.
67. Li, X. HMM based action recognition using oriented histograms of optical flow field. *Electron. Lett.* **2007**, *43*, 560–561. [CrossRef]
68. Lu, W.L.; Little, J.J. Simultaneous tracking and action recognition using the PCA-HOG descriptor. In Proceedings of the 3rd Canadian Conference on Computer and Robot Vision (CRV'06), Quebec City, QC, Canada, 7–9 June 2006; p. 6.
69. Thurau, C. Behavior histograms for action recognition and human detection. In *Human Motion–Understanding, Modeling, Capture and Animation*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 299–312.
70. Santiago-Mozos, R.; Leiva-Murillo, J.M.; Pérez-Cruz, F.; Artes-Rodriguez, A. Supervised-PCA and SVM classifiers for object detection in infrared images. In Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, Washington, DC, USA, 21–22 July 2003; pp. 122–127.
71. Chang, C.C.; Lin, C.J. LIBSVM: a library for support vector machines. *Acm Trans. Intell. Syst. Technol. TIST* **2011**, *2*, 27. [CrossRef]
72. Vishwanathan, S.; Smola, A.J.; Vidal, R. Binet–Cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. *Int. J. Comput. Vis.* **2007**, *73*, 95–119. [CrossRef]
73. Schölkopf, B.; Smola, A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2002.
74. Lucas, B.D.; Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision. 1981. Available online: https://www.researchgate.net/publication/215458777_An_Iterative_Image_Registration_Technique_with_an_Application_to_Stereo_Vision_IJCAI (accessed on 15 July 2019).
75. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
76. Wang, H.; Schmid, C. Action Recognition with Improved Trajectories. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013.
77. Bay, H.; Tuytelaars, T.; Van Gool, L. Surf: Speeded up robust features. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; Springer: Berlin/Heidelberg, Germany; pp. 404–417.
78. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In Proceedings of the Scandinavian Conference on Image Analysis, Halmstad, Sweden, 29 June–2 July 2003; Springer: Berlin/Heidelberg, Germany; pp. 363–370.
79. Prest, A.; Schmid, C.; Ferrari, V. Weakly supervised learning of interactions between humans and objects. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 601–614. [CrossRef] [PubMed]
80. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [CrossRef] [PubMed]
81. Akpinar, S.; Alpaslan, F.N. Video action recognition using an optical flow based representation. In Proceedings of theIPCV'14—The 2014 International Conference on Image Processing, Computer Vision, and Pattern Recognition, Las Vegas, NV, USA, 21–24 July 2014; p. 1.
82. Shi, J.; Tomasi, C. *Good Features to Track*; Technical Report; Cornell University: Ithaca, NY, USA, 1993.
83. Efros, A.A.; Berg, A.C.; Mori, G.; Malik, J. Recognizing action at a distance. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; p. 726.
84. Tran, D.; Sorokin, A. Human activity recognition with metric learning. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 548–561.
85. Ercis, F. Comparison of Histogram of Oriented Optical Flow Based Action Recognition Methods. Ph.D. Thesis, Middle East Technical University, Ankara, Turkey, 2012.
86. Li, H.; Achim, A.; Bull, D.R. GMM-based efficient foreground detection with adaptive region update. In Proceedings of the 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, 7–10 November 2009; pp. 3181–3184.
87. Sehgal, S. Human Activity Recognition Using BPNN Classifier on HOG Features. In Proceedings of the 2018 International Conference on Intelligent Circuits and Systems (ICICS), Phagwara, India, 19–20 April 2018; pp. 286–289.

88. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011.
89. 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.
90. Marszałek, M.; Laptev, I.; Schmid, C. Actions in context. In Proceedings of the CVPR 2009-IEEE Conference on Computer Vision & Pattern Recognition, Miami Beach, FL, USA, 20–25 June 2009; pp. 2929–2936.
91. Niebles, J.C.; Chen, C.W.; Fei-Fei, L. Modeling temporal structure of decomposable motion segments for activity classification. In Proceedings of the European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 392–405.
92. Zhang, Z. Microsoft kinect sensor and its effect. *IEEE Multimed.* **2012**, *19*, 4–10. [CrossRef]
93. Keselman, L.; Iselin Woodfill, J.; Grunnet-Jepsen, A.; Bhowmik, A. Intel realsense stereoscopic depth cameras. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 1–10.
94. Chen, J.; Wang, B.; Zeng, H.; Cai, C.; Ma, K.K. Sum-of-gradient based fast intra coding in 3D-HEVC for depth map sequence (SOG-FDIC). *J. Vis. Commun. Image Represent.* **2017**, *48*, 329–339. [CrossRef]
95. Liang, B.; Zheng, L. A survey on human action recognition using depth sensors. In Proceedings of the 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Adelaide, SA, Australia, 23–25 November 2015; pp. 1–8.
96. Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time human action recognition based on depth motion maps. *J. -Real-Time Image Process.* **2016**, *12*, 155–163. [CrossRef]
97. El Madany, N.E.D.; He, Y.; Guan, L. Human action recognition via multiview discriminative analysis of canonical correlations. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA , 5–28 September 2016; pp. 4170–4174.
98. Yang, X.; Zhang, C.; Tian, Y. Recognizing actions using depth motion maps-based histograms of oriented gradients. In Proceedings of the 20th ACM international conference on Multimedia, Nara, Japan, 29 October–2 November 2012; ACM: New York, NY, USA, 2012; pp. 1057–1060.
99. Oreifej, O.; Liu, Z. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 716–723.
100. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Mining actionlet ensemble for action recognition with depth cameras. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1290–1297.
101. Wang, J.; Liu, Z.; Chorowski, J.; Chen, Z.; Wu, Y. Robust 3D action recognition with random occupancy patterns. In *Computer Vision–ECCV 2012*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 872–885.
102. Liu, M.; Liu, H.; Chen, C. Robust 3D action recognition through sampling local appearances and global distributions. *IEEE Trans. Multimed.* **2018**, *20*, 1932–1947. [CrossRef]
103. Seo, H.J.; Milanfar, P. Action recognition from one example. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 867–882. [PubMed]
104. Satyamurthi, S.; Tian, J.; Chua, M.C.H. Action recognition using multi-directional projected depth motion maps. *J. Ambient. Intell. Humaniz. Comput.* **2018**, 1–7. [CrossRef]
105. Ojala, T.; Pietikäinen, M.; Mäenpää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, 971–987. [CrossRef]
106. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]
107. Li, W.; Zhang, Z.; Liu, Z. Action recognition based on a bag of 3D points. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 9–14.
108. Kurakin, A.; Zhang, Z.; Liu, Z. A real time system for dynamic hand gesture recognition with a depth sensor. In Proceedings of the 20th European signal processing conference (EUSIPCO), Bucharest, Romania, 27–31 August 2012; pp. 1975–1979.

109. Xia, L.; Chen, C.C.; Aggarwal, J.K. View invariant human action recognition using histograms of 3D joints. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 20–27.
110. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale Video Classification with Convolutional Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
111. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
112. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
113. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1915–1929. [CrossRef]
114. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
115. Sharif Razavian, A.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN features off-the-shelf: An astounding baseline for recognition. *arXiv* **2014**, arXiv:1403.6382.
116. Dean, J.; Corrado, G.; Monga, R.; Chen, K.; Devin, M.; Mao, M.; Senior, A.; Tucker, P.; Yang, K.; Le, Q.V.; et al. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2012; pp. 1223–1231.
117. 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; pp. 568–576.
118. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
119. Crammer, K.; Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.* **2001**, *2*, 265–292.
120. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
121. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
122. Zaremba, W.; Sutskever, I. Learning to execute. *arXiv* **2014**, arXiv:1410.4615.
123. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y. Towards good practices for very deep two-stream convNets. *arXiv* **2015**, arXiv:1507.02159.
124. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
125. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
126. Wang, L.; Qiao, Y.; Tang, X. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4305–4314.
127. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
128. 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]
129. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941.

130. 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, 7–12 June 2015; pp. 4694–4702.
131. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 20–36.
132. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
133. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
134. 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, 7–13 December 2015; pp. 4597–4605.
135. Bilen, H.; Fernando, B.; Gavves, E.; Vedaldi, A.; Gould, S. Dynamic image networks for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3034–3042.
136. Fernando, B.; Gavves, E.; Oramas, J.M.; Ghodrati, A.; Tuytelaars, T. Modeling video evolution for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5378–5387.
137. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; ACM: New York, NY, USA, 2014; pp. 675–678.
138. Carreira, J.; Zisserman, A. Quo vadis, action recognition? A new model and the Kinetics dataset. In Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4724–4733.
139. Varol, G.; Laptev, I.; Schmid, C. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1510–1517. [CrossRef]
140. Taylor, G.W.; Fergus, R.; LeCun, Y.; Bregler, C. Convolutional learning of spatio-temporal features. In Proceedings of the European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; pp. 140–153.
141. Ullah, A.; Ahmad, J.; Muhammad, K.; Sajjad, M.; Baik, S.W. Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features. *IEEE Access* **2018**, *6*, 1155–1166. [CrossRef]
142. Graves, A.; Fernández, S.; Schmidhuber, J. Bidirectional LSTM networks for improved phoneme classification and recognition. In Proceedings of the International Conference on Artificial Neural Networks, Warsaw, Poland, 11–15 September 2005; pp. 799–804.
143. Wang, J.; Cherian, A.; Porikli, F.; Gould, S. Video representation learning using discriminative pooling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1149–1158.
144. Schindler, K.; Van Gool, L. Action snippets: How many frames does human action recognition require? In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AK, USA, 24–26 June 2008.
145. Wang, X.; Gao, L.; Wang, P.; Sun, X.; Liu, X. Two-stream 3D convNet fusion for action recognition in videos with arbitrary size and length. *IEEE Trans. Multimed.* **2018**, *20*, 634–644. [CrossRef]
146. Liu, J.; Luo, J.; Shah, M. Recognizing realistic actions from videos in the wild. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
147. Wang, X.; Farhadi, A.; Gupta, A. Actions$\sim$ transformations. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2658–2667.
148. Chaquet, J.M.; Carmona, E.J.; Fernández-Caballero, A. A survey of video datasets for human action and activity recognition. *Comput. Vis. Image Underst.* **2013**, *117*, 633–659. [CrossRef]
149. UCF101. Action Recognition Data Set. Available online: https://www.crcv.ucf.edu/data/UCF101.php (accessed on 15 July 2019).

150. UCF50. Action Recognition Data Set. Available online: https://www.crcv.ucf.edu/data/UCF50.php (accessed on 15 July 2019).
151. HMDB: A large human motion database. Available online: http://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/ (accessed on 15 July 2019).
152. Actions as Space-Time Shapes. Available online: http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html (accessed on 15 July 2019).
153. MSR Action Recognition Dataset. Available online: http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/ (accessed on 15 July 2019).
154. 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, 7–12 June 2015; pp. 961–970.
155. A Large-Scale Video Benchmark for Human Activity Understanding. Available online: http://activity-net.org/ (accessed on 15 July 2019).
156. Goyal, R.; Kahou, S.E.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; Volume 1, p. 3.
157. The 20BN-something-something Dataset V2. Available online: https://20bn.com/datasets/something-something (accessed on 15 July 2019).
158. The Sports-1M Dataset. Available online: https://github.com/gtoderici/sports-1m-dataset/blob/wiki/ProjectHome.md (accessed on 15 July 2019).
159. YouTube-8M: A Large and Diverse Labeled Video Dataset for Video Understanding Research. Available online: https://research.google.com/youtube8m/ (accessed on 15 July 2019).
160. Gu, C.; Sun, C.; Ross, D.A.; Vondrick, C.; Pantofaru, C.; Li, Y.; Vijayanarasimhan, S.; Toderici, G.; Ricco, S.; Sukthankar, R.; et al. AVA: A video dataset of spatio-temporally localized atomic visual actions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6047–6056.
161. AVA: A Video Dataset of Atomic Visual Action. Available online: https://research.google.com/ava/explore.html (accessed on 15 July 2019).
162. 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, 7–12 June 2015; pp. 204–212.
163. A Universal Labeling Tool: Sloth. Available online: https://cvhci.anthropomatik.kit.edu/~baeuml/projects/a-universal-labeling-tool-for-computer-vision-sloth/ (accessed on 15 July 2019).
164. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [CrossRef]
165. LabelMe. Available online: http://labelme.csail.mit.edu/Release3.0/ (accessed on 15 July 2019).
166. LabelBox. Available online: https://labelbox.com/ (accessed on 15 July 2019).