sensors

MDPI

Review

A Survey of the Techniques for The Identification and Classification of Human Actions from Visual Data

Shahela Saif *, Samabia Tehseen and Sumaira Kausar

Computer Science Department, Bahria University, E-8 Islamabad 44000, Pakistan;
stehseen.buic@bahria.edu.pk (S.T.); sumairakausar@bui.edu.pk (S.K.)
* Correspondence: shehlasaif7@gmail.com; Tel.: +92-51-9049-5367

Received: 29 September 2018; Accepted: 9 November 2018; Published: 15 November 2018

Abstract: Recognition of human actions form videos has been an active area of research because it has applications in various domains. The results of work in this field are used in video surveillance, automatic video labeling and human-computer interaction, among others. Any advancements in this field are tied to advances in the interrelated fields of object recognition, spatio- temporal video analysis and semantic segmentation. Activity recognition is a challenging task since it faces many problems such as occlusion, view point variation, background differences and clutter and illumination variations. Scientific achievements in the field have been numerous and rapid as the applications are far reaching. In this survey, we cover the growth of the field from the earliest solutions, where handcrafted features were used, to later deep learning approaches that use millions of images and videos to learn features automatically. By this discussion, we intend to highlight the major breakthroughs and the directions the future research might take while benefiting from the state-of-the-art methods.

Keywords: computer vision; action recognition; visual action recognition; deep learning

1. Introduction

Activity recognition involves an understanding of human actions. A human action is harder to define than to understand, and many attempts have been made in the literature to define it in one way or the other. Turaga et al. [1] provided an intuitive definition of an action as “simple motion patterns usually executed by a single person and typically lasting for a very short duration (order of tens of seconds)”. Moeslund and Graum [2] and Poppe [3] have defined action as “an atomic movement that can be described at limb level.”; whereas the activity can be considered a sequence of actions that can involve interactions among humans or between humans and the environment.

The recognition of human actions form videos is a challenging task. It requires work in multiple disciplines to be effectively executed and combined such as object recognition, background and foreground processing, semantic segmentation and human dynamics. There are two major types of recognition systems: one that involves the use of wearable sensors or associated devices and the other that uses cameras and wireless radio frequency modules. Among the first kind, a few approaches to action detection have involved the use of dedicated sensors such as mobile sensors [4–6] or physiological data [7,8]. Classifiers are used on these data for action recognition. These approaches promise a higher accuracy, but work in a limited domain. In the second approach, features are extracted from visual input including single object’s features such as position, shape, color or global features such as region occupancy or positional variations. Normal activity templates and abnormal activity templates are created that can be subjected to recognition through template matching methods or state space [9]. In recent years, there has been a significant increase in the uses of multi-modal video devices such as Kinect, which provides depth information apart from the color information from (regular) video cameras. Such systems can provide an accurate representation of a human shape, which is utilized to form various activity shape features [10]. Researchers have used kinematic joints [11], human posture [12] and even histogram-based approaches [13] for action recognition using such devices [14]. Since our focus is on devices that use traditional video data that do not include depth information, we shall not discuss these any further in the current study.

Video analysis has been performed at various levels of detail depending on the information we require from them. The few significant ones were given in a study by [15]:

1. Object scope understanding where only the positions of persons and objects are detected.
2. Tracking scope understanding where the trajectories and correspondence of objects are analyzed.
3. Pose-level understanding that involves the analysis of the position of human body parts.
4. Analysis of human activities and events.

There are several existing surveys that have explored the techniques for activity recognition. Some of these have divided these recognition approaches into single-layered and hierarchical approaches, as in the works of Aggarwal and Ryoo [15] and Cheng et al. [16]; while others like Moeslund et al. [2] and Poppe [3] have divided the work on the basis of action and activity. Aggarwal and Cai [17] have performed another survey on the same domain in which they reviewed the literature from three perspectives: (1) motion analysis with regards to body parts, (2) tracking from single or multiple camera perspectives and (3) using images for recognizing activities. Gavrila [18] also discussed action recognition techniques based on whole-body or hand motion tracking while discussing both 2D and 3D approaches.

Handcrafted feature extraction techniques paired with classifiers have been used for action recognition for quite some time and with considerable success [19]. However, the availability of large amounts of data has made possible the use of deep networks for the task of action recognition [20]. The success of deep networks and in particular CNN is evident from the results on ImageNet [21]. A mention of the other studies that cover action recognition is provided in Table 1. This survey is oriented in a manner to review both the handcrafted recognition techniques and the deep learning techniques, as given in Figure 1. We also explore the effect of using local features for action recognition. Figure 2 shows the research interest in action recognition over the years. With time, the handcrafted approaches matured and started producing results that could be used in building real-time applications. The renewed interest came with the arrival of deep architectures in 2012 and later. There are many studies that have explored the applicability of deep architectures to activity recognition, both in conjunction with handcrafted approaches and in standalone capacity.

Table 1. Surveys and studies on action and motion analysis.

Survey	Scope
Poppe [4]	Handcrafted action features and classification models
Aggarwal and Ryoo [15]	Individual and group activity analysis
Turaga et al. [1]	Human actions, complex activities
Moeslund et al. [2]	Human action analysis
Poppe [3]	Human action recognition
Cheng et al. [16]	Handcrafted models
Aggarwal and Cai [17]	Human action analysis
Gavrila [18]	Human body and hands tracking-based motion analysis
Yilmaz et al. [22]	Object detection and tracking
Zhan et al. [23]	Surveillance and crowd analysis
Weinland et al. [24]	Action recognition
Aggarwal [25]	Motion analysis fundamentals
Chaaraoui et al. [26]	Human behavior analysis and understanding
Metaxas and Zhang [27]	Human gestures to group activities
Vishwakarma and Agrawal [28]	Activity recognition and monitoring
Cedras and Shah [29]	Motion-based recognition approaches

The rest of the paper is organized as follows: In the next section, we discuss some of the challenges of action recognition using data from videos. Section 3 gives an overview of the handcrafted approaches that essentially use handcrafted methods for identification of action in conjunction with a classifier for action classification. In Section 4, we take a look at the approaches that use deep learning. The deep learning approaches include: (i) approaches that make use of handcrafted features for identification that are given to a deep network for fine-tuning and classification; (ii) approaches that use deep networks both for the task of feature extraction and classification; (iii) hybrid approaches; and (iv) deep generative models. A critical discussion of the approaches follows the details of the datasets prior to the conclusion.

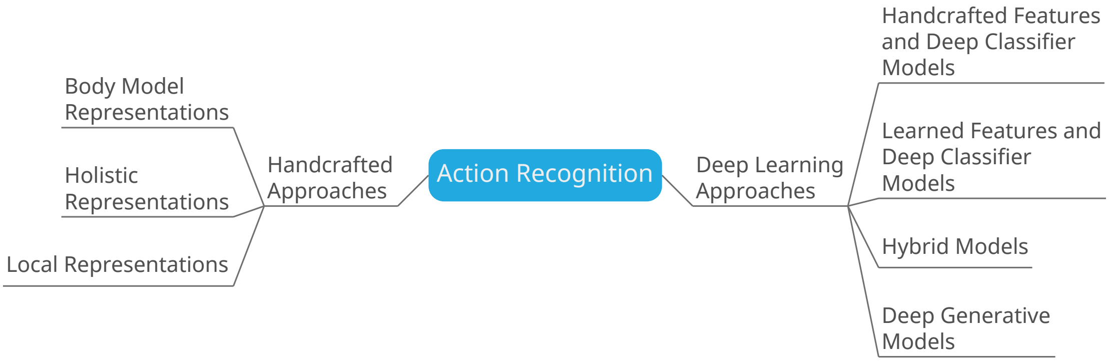

Figure 1. Classification of action recognition based on techniques employed for identification and classification of actions.

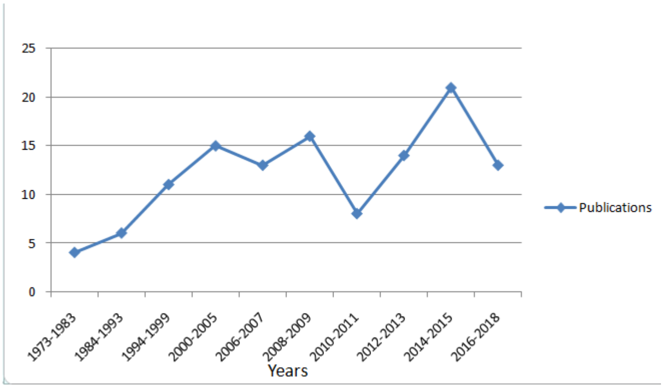

Figure 2. Research publications per year as discussed in the current study.

2. Challenges

The activity recognition process involves quite a few challenges and constraints that need to be dealt with at the time of both feature extraction and classification. Poppe has listed some of the significant ones in [3]. These are:

1. Inter-class variations: Different people perform different actions in their own ways, which at times show very low resemblance to one another, e.g., walking methods may differ in stride length or speed.
2. Intra-class similarities: Actions belonging to different classes may appear similar such as jogging and running.
3. View point variations: The same action if observed from two independent viewpoints can appear to be different, and the data collected as a result may indicate separate classes.
4. Environment: Cluttered or complex backgrounds can make the task of identification of clear human shapes much more difficult.

5. Temporal variations: Temporal variations occur both in terms of action performance/completion and action observation.

All these issues are addressed explicitly by the action recognition approaches as and when they arise. However, depending on the datasets that are being used and the feature selection techniques employed, the impact of these constraints may vary. There is, thus, no single strategy that can be applied for any particular problem while using different action recognition techniques. In the subsequent sections, we provide a review of various action recognition techniques along with their shortcomings. The organization of the techniques is based on the time of introduction and the growing complexity of the presented techniques.

3. Handcrafted Approaches

The interest in human action recognition is not a recent one, and scientists and researchers have over time been utilizing various techniques for action identification. Using spatial information about the human pose, which is generated by extracting various image features, we can classify human pose based on the similarity of the pose to some action.

3.1. Body Models

Among the earliest attempts at action recognition, Johansson [30] used a simplistic representation of the human body that was comprised of readable light sources placed on joints (Moving Light Displays (MLDs)) and could determine the action based on the movement of joints. An example of these MLDs is given in Figure 3.

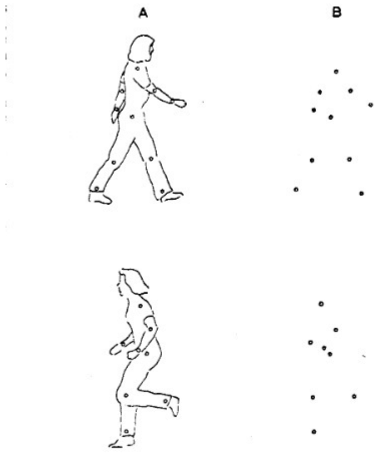

Figure 3. Moving light displays used for action recognition in [30].

As a pioneer work in this field, these simplistic experiments paved the way for many more methods based on the same idea. The two predominant techniques that emerged as a consequence of this work are the (i) representation of motion as a 2D sequence of actions and (ii) generation of 3D models from 2D representations to recognize actions [24]. The variability of the human body's shape poses many interesting challenges that have led researchers to construct 3D models of the human body. The earliest work along these lines was done by Marr and Nishihara [31], where they used cylindrical models for human body representation, as given in Figure 4.

Others have built on such models as well [32,33]; some have provided more flexible models using super quadrics [34] and textured spline models [35]. These models are difficult to compute and do not have the flexibility to provide solutions to problems such as view point variations, environment clutter or temporal variations. These worked in strictly controlled environments and were therefore soon replaced by improved techniques. They did, however, set the direction for future research for many years to come. The concepts of body models were picked up by researchers who used wearable devices and 3D data-collection devices such as Kinect for action recognition [4]. Such models have also provided accuracies up to 90%. These are not discussed in detail here, as the aim of the current study is 'visual data'-only techniques.

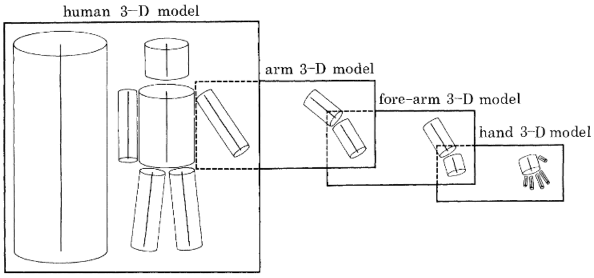

Figure 4. Human model created in 3D using 2D information in [31].

3.2. Holistic Representations

Holistic representations do not require identification or marking of individual body parts unlike the body models discussed in the previous section. These approaches work by preprocessing the images by performing fundamental tasks such as background subtraction and feature extraction. Most techniques make use of contours and/or silhouettes of the human body [36–38].

Darell and Pentland [39] created a model where images of hand gestures were correlated with one another directly without the need to extract any features. However, for their work, they assumed a static black background, which may not be very practical. A significant work in the same direction was by Yamato et al. [40] in which they converted the time-sequential images into a unified image feature vector where only silhouettes were used. This feature vector is used as a symbol sequence that is evaluated using a Hidden Markov Model (HMM).

Work by Bobick and Davis [41] has had a tremendous effect on all future research on activity recognition. They created 'Motion History Images (MHI)' and 'Motion Energy Images (MEI)' from silhouettes that were integrated over the time domain (using frames' information); see Figure 5 for a reference. MHI and MEI have been adapted and improved by many later works. Space-time volumes that a silhouette spans over in multiple frames were used in [42,43] as opposed to integrating the time-sequence into one image, as done by [41].

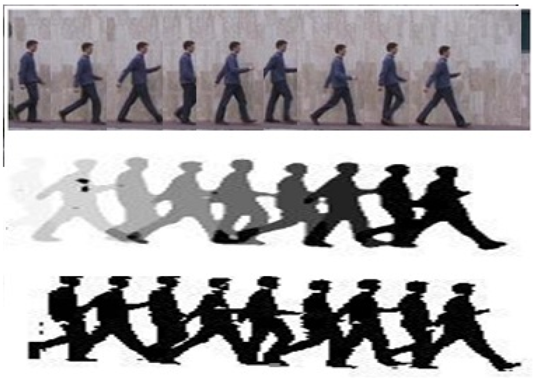

Figure 5. Top row: A walking sequence of a person; middle row: a Motion Energy Image (MEI) template; bottom row: a Motion History Image (MHI) template [41].

Elgammal et al. [44] and Weinland and Boyer [45] used chamfer distance to eliminate the affects of noisy silhouettes, which are caused due to cluttered backgrounds. Shape-context descriptors [46–48] were also used to the same effect. Silhouettes are insensitive to color, texture and context, but are not very effective in cases of self-occlusion. A better approach is the use of dense optical flows [49,50] and clustering these optical flows into motion blobs [51]. Optical flow fields were split into four different scalar fields by [52–54]. Optical flow fields do not require background subtraction, but are also sensitive to material properties, lightening, etc. Gradients are also used to extract image features [55]. Histograms of oriented gradients are used for object detection [56] and for action recognition [57]. Gradient features, like optical flows, do not require background subtraction, but are affected by material properties. Some studies have used optical flows in combination with gradients [58,59] and silhouettes [60] to achieve superior results.

The results of holistic representations are promising, but are incapable of handling viewpoint variations [51,61]. Improvements of these are local and deep approaches.

3.3. Local Representations

3.3.1. Interest Point Detection

Work by Laptev [62] on space-time interest points paved the way for local representations for image feature extractions. The author adapted the Harris corner detector [63] to create a 3D-Harris detector that can detect spatial changes in orthogonal directions along with points that have large non-constant motion, as seen in Figure 6. The 3D-Hessian detector [64] uses second order derivatives instead of gradients as in the Harris detector for interest point detection.

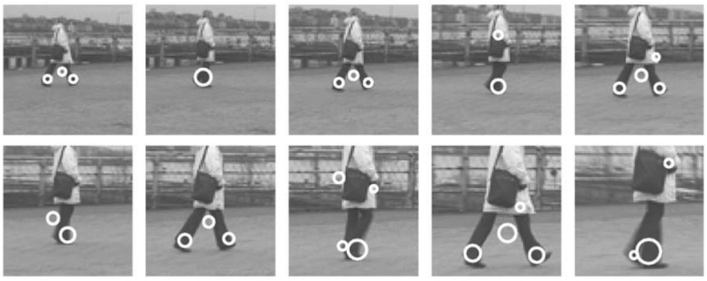

Figure 6. Spatio-temporal interest point detection for a walking person. Reprinted with permission from [62].

3.3.2. Local Descriptors

Earlier works in action recognition have used cuboid models for body representation [62,65], but were challenged by Messing [66] and Matikainen et al. [67] in terms of effectiveness and flexibility. An improvement on this is considered by using edge and motion descriptors.

Edge and Motion Descriptors

Histogram of oriented Gradients (HoGs) were used for motion detection by [68], and later, [56] extended this to the spatio-temporal domain, naming it HoG3D. Laptev [58] employed the same idea for optical flow fields, since they encode the pixel-level motion in videos, and created the Histogram of optical Flow (HoF). Dalal et al. [69] created a more robust version of HoF, the Motion Boundary Histogram (MBH). The calculation of optical flow fields is computationally expensive, and decompression techniques have been employed [70] to overcome this disadvantage.

3.3.3. Trajectory-Based Approaches

One criticism to cuboid representations is that over a span of frames, the detected interest point may not lie at the same spatial location within the temporal bounds of a cuboid. Action trajectory is the tracking of a feature in the time domain. Trajectory-based action representations were widely adopted after the works of Messing et al. [66] and Matikainen et al. [67]. Wang et al. [71] integrated MBH, HoG and HoF to create a rich feature representation, where trajectories were calculated by using optical flow.

Vig et al. [72], addressing the computational complexity of the prior technique, used 'saliency-maps' to extract the region of interest inside frames. In a similar approach, Jiang et al. [73] used local and global reference points along with trajectories to improve motion detection. Wang et al. [74] improved their original work by eliminating the effect of camera movements by using SURF and dense optical flows. The improved model was adopted by many, including Peng et al. [75], who have developed a multi-layer stacked Fisher Vector (FV) [76] with improved performance over the original model.

Handcrafted approaches are complex to build and hard to modify. These cannot be readily adapted to new or complex datasets, which has hindered their ability to provide a unified global solution. This was changed by the rapid increase in use of deep architectures for image analysis techniques. Given in the next section are various approaches based on deep learning; some of which also make use of handcrafted approaches in a limited capacity.

4. Deep Learning Approaches

With the advent of deep learning approaches that enable the learning of features along with the classification of them, we have seen the application of these in the field of action recognition with considerable success. In particular, convolutional neural networks have revolutionized the field of image classification and recognition [77–80] and are employed singularly or in conjunction with other architectures for action recognition tasks.

In general, we can categorize deep approaches into two major schemes based on network function: supervised approaches and unsupervised approaches. The supervised approaches include (i) networks that extract features from deep models and use other classifiers and (ii) networks that use deep models for end-to-end classification, as well as (iii) networks that use handcrafted features in conjunction with deep networks for classification [81]; while unsupervised and semi-supervised approaches are the deep generative models, such as autoencoders or adversarial networks.

The supervised approaches are split into three architectures or combinations and/or evolutions of these three architectures: Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and Long Short-Term Memory networks (LSTM).

CNN: CNN consist of a number of convolutional layers, each of which is responsible for feature extraction. The lower layers extract simple features, while the higher layers extract complex features by the use of filters at each layer. The filters are designed on the principle of weight sharing, which enables reducing the number of parameters to learn. Each layer that increases the depth and complexity of a network also inadvertently increases the dimensionality of the convolved features. CNN are used as effective feature learners, but their greatest strength is their ability to be used as end-to-end models for classification [82].

RNN: Recurrent neural networks have the ability to process feedback connections, which allows them to model sequential behavior. RNNs have found considerable success in handwriting recognition [83,84] and speech recognition [85,86], which led to their induction to modeling temporal associations among video frames to represent human action. The recurrent neural network effectively updates its current memory vector depending on three elements: current frame, previous memory vector and previous location of an object.

LSTM: Long short-term memory models are used in conjunction with various CNN and/or RNN models in order to represent long-term temporal dynamics and to do away with the vanishing gradient problem.

4.1. Handcrafted Features and Deep Classifiers

Handcrafted features have given promising results over the span of decades, where more and more sophisticated features emerged with time [71–75]. The appeal of using handcrafted features is to incorporate the time dimension of video sequences and to provide a 'running start' to a deep network. Kim et al. [87] proposed a modified convolutional neural network where the low level action information is represented by handcrafted features. The action sequence of any person in a video generates a 3D volume that is extracted using 3D Gabor filters [88]. These filters extract the outer boundary of an actor in a 2D or spatial plane, and when considered across multiple frames, they generate a spatio-temporal volume. These spatio-temporal volumes make the actions view-invariant. A 3D CNN is applied to each spatio-temporal volume, and features are extracted based on these. The features thus obtained are classified using a discriminative classification model [87]. Jhuang et al. [89] created a feed-forward hierarchical framework that detects spatio-temporal features of increasing complexity to measure 'motion-direction sensitive units'. By taking the global max of each feature map containing scale- and position-invariant features [90–92], a feature vector is computed as a final representation. The approach is sensitive to the effectiveness of the handcrafted spatio-temporal feature detectors, which limits its effectiveness.

4.2. Learned Representations and Deep Classifiers

The three-dimensional convolutional neural networks aim to extract spatial features using the normal 2D transforms and employ the third dimension to extract temporal information [93]. The 3D convolutional network as presented by Ji et al. in [93] applies a 3D kernel, a spatial kernel extended in the time dimension by applying the same 2D filter to a particular spatial location in multiple frames. This makes the features obtained by 3D convolutions invariant to spatial translation with respect to time. The 3D convolutional neural network is shown to produce better results than its 2D counterparts [93]. Most 3D architectures constructed in this manner have a limit to the number of frames used for extracting information in the temporal domain, which makes them very rigid. However, a major restriction is the high computational cost and need for a large amount of trained data. Varol et al. [94] used longer temporal regions for performing 3D convolutions, and it was seen that extending temporal depth improves the performance of the network.

Research has been focused on how to successfully incorporate the time dimension in deep networks. Ng et al. [95] have worked on the idea of temporal pooling and showed that max pooling provides the best results. Karpathy et al. [96] created different models for combining information from spatial and temporal domains; early fusion, late fusion and slow fusion; see Figure 7 for a reference. The single-frame approach uses one frame and applies a deep architecture over it without using temporal information. In late fusion, two images a certain number of frames apart are fed to two independent networks, fusing the results at the fully connected (FC) layers. Early fusion merges the frames at the pixel level before running them through a network. Slow fusion is an amalgam of late fusion and early fusion. The procedure requires the convolutional layers to be connected across multiple frames, thus providing the benefit of temporal convolution along with spatial convolution. Among the three, slow fusion performs better than the others because of its use of 3D convolutional kernels across multiple layers. Karpathy et al. have also experimented with multi-resolution models by creating a two-stream network. The 'context' stream processes a low resolution complete image, and a 'fovea' stream processes a high resolution cropped center of the image. The results of convolutions on both streams are combined at fully-connected layers to produce classification results. Using multi-resolution videos in separate, but identical networks reduces the number of parameters to learn and improves the accuracy [96]. Tran et al. [97] also used 3D CovNets while making use of a small 3 × 3 × 3 convolutional kernel throughout the network and showed that constant depth at every layer performed better than varying the temporal depth at each layer. This network, named C3D, gives rise to a generic descriptor, that averages the outputs of the fully-connected layers, with the aim of learning generic features from video, so that the network would not have to be fine-tuned for each independent task [97].

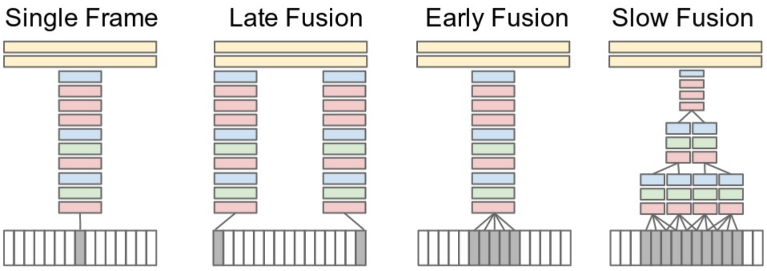

Figure 7. Fusion strategies for incorporating the temporal dimension in neural networks. Source: Reprinted with permission from [96].

Using 3D filters increases the number of parameters and inevitably increases the cost and complexity of the network. Sun et al. [98] addressed this issue in their work and suggested factorizing a 3D filter into a 2D filter and a 1D filter. The benefit is reducing the number of network parameters from $n_x n_y n_t$ to $n_x n_y + n_t$, thus reducing the problem of kernel complexity by a factor of $O(n_t)$. Others have exploited recurrent structures for achieving the same goals. Baccouche et al. [99] and Donahue et al. [100] have used a cascade of convolutional neural networks with Long-Short Term Memory (LSTM), where LSTMs are a class of recurrent networks [101]. In the work by Donahue et al. [100], the network named the Long-term Recurrent Convolutional Network (LRCN) performed an end-to end-training. The model has been successfully used not only for action recognition, but also for captioning of images and videos.

4.3. Hybrid Models

Multi-stream models have been built on the idea of the separation between the spatial domain and temporal domain. Simonyan et al. [102] introduced this idea of multiple streams where they trained one convolutional network to extract spatial information about the video frames and another to capture temporal information using optical flows [103]. The two streams were later 'fused' using their softmax rates, as shown in Figure 8. They [102] have worked with layers of dense optical flows of consecutive frames, motion trajectories and bi-directional optical flows. The convolutional neural network is trained in a multi-task learning setting [104] by classifying both on the HMDB-51 and UCF-101 datasets and using two softmax layers, each of which computes a score on its respective dataset. Both streams are trained in the same manner, where the temporal network is an adaptation of the model by [105].Fiechtenhofer et al. [106] have shown improved results with a similar architecture that performs fusion at an intermediate layer.

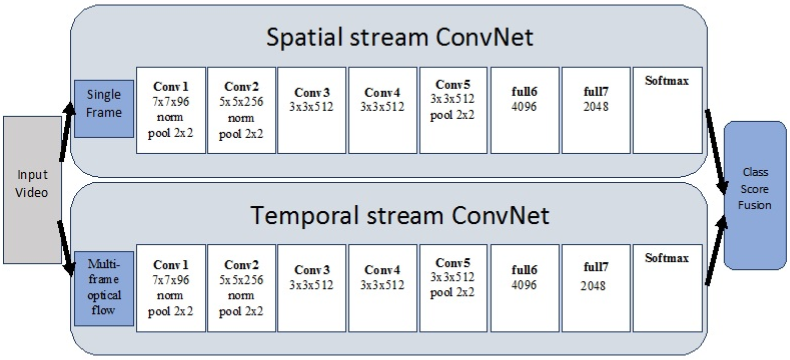

Figure 8. Two-stream architecture with the spatial stream using images and the temporal stream using optical flows. Source: [102].

Other works that have explored the idea of multi-stream networks are [107] in the form of a trajectory of pooled two-stream deep convolution descriptors. The network architecture is similar to that of [102], and they have used UCF-101 and HDMB-51 to compute multi-scale feature maps of the videos. Wang et al. [107] have further aggregated the computed dense trajectories over feature maps using the Fisher vector [108], but in terms of performance, this approach is no better than the original two-stream network.

4.4. Deep Generative Models

There is an ever-increasing amount of video data that are available over the Internet, but since most of this is consumer generated, thus they are not labeled. The potential of being able to use these data in an unsupervised environment to understand and predict the action sequences can give rise to endless possibilities. Generative models built for sequence analysis [79,109] have the ability to predict the next state of a sequence x_{t+1} given a sequence of states $S = x_1, x_2, ...x_t$. Deep generative models do not require labels for training, but rely on finding accurate motion primitives [110–112]. Autoencoders have been used in research by many [113–115] for unsupervised learning of features through deep architectures. Yan et al. [116] captured video dynamics using a deep autoencoders, Dynencoder. The first layer of the model maps the input to hidden states; the second or prediction layer predicts the next hidden states based on the current ones, and the final layer is from the predicted hidden states to estimated input states. The training phase is followed by end-to-end fine-tuning.

Sirvastava et al. [109] created an LSTM autoencoder where two recurrent structures were used: encoder LSTM and decoder LSTM. The encoder LSTM receives input and learns compact representations, while the decoder LSTM uses these learned representations to reconstruct the input. An LSTM autoencoder can be used to predict the next states of a sequence, as well, and is thus more efficient than a 3D CNN. Another interesting approach is the use of adversarial networks [110]. In this work, two models were trained simultaneously; (i) a generative model that takes input data and generates a representation of them; and (ii) a discriminative model that tries to distinguish between real input and a generated representation. The harder it is for the discriminative model to differentiate between real and generated data, the better the learned representation and, thus, the model performance [110]. Mathieu et al. [117] have also used the adversarial model to train multi-scale CNN that avoid pooling layers. Their model is for video prediction, as well.

5. Datasets

The datasets for action recognition have evolved over time and have become more complex and realistic [20]. The earliest datasets such as KTH and Weizmann have a fixed number of subjects and a very limited number of action categories, as they were shot in controlled environments [118]. Datasets with increasing complexity include not only more action classes, but also complex backgrounds, multiple actors, occlusions and viewpoint variations; some even contain resolution inconsistencies [118]. A list of datasets used for action recognition is given in Table 2.

Table 2. Datasets used for action recognition in increasing order of complexity.

Dataset	Type	No. of Videos	No. of Classes	No. of Subjects
KTH [119]	Indoor/Outdoor	600	6	25
Weizmann [42]	Outdoor	90	10	9
CAVIAR [120]	Indoor/Outdoor	80	9	numerous
UCFSports [121]	Television sports	150	10	numerous
UCF-50 [122]	YouTube videos	-	50	numerous
UCF-101 [123]	YouTube videos	13,320	101	numerous
Sports-1 M [96]	YouTube sports	1,133,158	487	numerous
Hollywood2 [124]	Clips from Hollywood movies	1707	12	numerous
HMDB-51 [125]	YouTube, movies	7000	51	numerous

The most challenging datasets are the ones that involve YouTube videos and sports videos. These have the most variable backgrounds and viewpoint variations. Some YouTube videos are from user devices and have low camera stability and low resolution. A list of techniques and their accuracy is presented in Table 3 for further discussion. Nearly all of these papers have reported results on more than one dataset, but we have chosen to show only the ones that have reported the highest accuracy.

Table 3. Comparison of various action recognition techniques.

Paper	Year	Technique	UCF-101	HMDB-51	Others
Handcrafted Features					
Wang et al. [71]	2011	Dense Trajectory			UCF Sports 88.2
Wang et al. [74]	2013	Dense Trajectory			UCF-50 91.2
Learned Models					
Ji et al. [93]	2013	3D Convolution			KTH 90.2
Tran et al. [97]	2015	C3D generic descriptor	90.4		
Karpathy et al. [96]	2014	Slow fusion			Sports-1 80.2
Sun et al. [98]	2015	Factorized spatiotemporal CovNets	88.1	59.1	
Wang et al. [107]	2015	Two-stream	89.3		
Ng et al. [95]	2015	Conv Pooling		88.2	Sports-1 73.1
Ng et al. [95]	2015	LSTM		88.6	
Donahue et al. [100]	2015	LRCN	82		
Jiang et al. [73]	2012	Trajectories	78.5	48.4	
Varol et al. [94]	2017	Long-term temporal convolutions	91.7	64.8	
Li et al. [126]	2016	VLAD	92.2		
Hybrid Models					
Simonyan and Zisserman [102]	2014	Two-stream CNN	88.0	59.4	
Feichtenhofer et al. [106]	2016	ResNet	93.5	69.2	
Wang et al. [107]	2015	Trajectory pooling + Fisher vector	91.5	65.9	
Lev et al. [127]	2016	RNN Fisher vector	94.08	67.71	
Bilen et al. [128]	2016	Dynamic Image network	89.1	65.2	
Wu et al. [129]	2015	Adaptive multi-stream fusion	92.6		
Deep Generative Models					
Srivastava et al. [109]	2015	LSTM autoencoder	75.8	44.1	
Mathieu [117]	2015	Adversarial network	≈ 90		

6. Discussion

It is interesting to see how the deep approaches in action recognition perform with respect to handcrafted or local approaches; since in terms of images, we have seen that deep architectures have outperformed the previous approaches by quite a wide margin [105]. An accurate comparison of the performance of the models can only be done after taking into consideration the datasets they have used. The deep networks have not shown the same amount of improvement over handcrafted feature techniques in video processing as they had in image processing. Some of the state-of-the-art handcrafted approaches are on par with deep approaches. Handcrafted approaches like 'dense trajectory' [71,74] have provided better results than some of the deep approaches, such as in [96,102], as is evident from Table 3. A possible reason might be that the available labeled images datasets are much larger than the labeled video datasets. Another consideration is that the architecture of CNN, which is the most widely used for image classification, is inherently better suited for treating images as independent elements and does not have the ability to directly incorporate time information spanning over multiple sequences. For this purpose, we have seen the use of RNNs and LSTMs to be able to add sequence-related information into models [95,100,127].

Even though much of the research has shifted towards adapting deep networks for action recognition tasks, deep networks have not completely replaced the traditional handcrafted approaches. A few approaches have focused on getting the benefit of both techniques, i.e., handcrafted features and learned representations, by employing the concept of 'transfer learning' as in the works of [96]. Dense trajectory solutions [74] are an example of how well the handcrafted approaches can perform on smaller, but challenging datasets, where deep approaches are limited by the size and quality of datasets. A majority of learning-based approaches to action recognition either directly apply CNN to videos or employ a variation of it to learn features.

In deep networks, spatio-temporal networks and two-stream networks have given better performance than their counterparts. Both of these solutions build on the traditional 3D convolutional architecture by using 3D filters. To obtain temporal information, dedicated streams that use optical flow trajectories have been used, which have been very successful on datasets [102,107], but have the problem of over-fitting. The flow trajectories trained on one set cannot be effective to the same degree on all sets. Deeper networks also perform better than shallower ones [107], but training deeper networks requires better augmented data available in larger amounts other than the severe resource constraint they apply in terms of the number of parameters to tune.

One area that will require further exploration in the future is the idea of pairing video recognition architectures with image recognition ones [20]. Furthermore, multi-stream networks that carry forward more context information should be explored in conjunction with spatial feature recognizers. LSTMs have also shown promising results [95], and their recurrent nature may support the transfer of more complex context information. It is yet to be seen how unsupervised and semi-supervised techniques can be used in conjunction with supervised ones to improve the overall results.

7. Conclusions

The ability of machines to understand images and scenes has driven many researchers to find incredible solutions by machine learning. We saw that from simple techniques like MLD (Moving Light Displays) to deep approaches, over time, many solutions have been proposed to find a solution to this problem. Techniques that were used for image understanding have been extended to work for action recognition through videos, as well, with considerable success. However, the problem of action recognition through videos is far more complicated than image analysis. A discussion has been presented to find the techniques that have been used over time and to highlight the most successful ones, in the two dominant categories of 'deep learning' approaches and 'non- deep learning' approaches, while finding the direction for future research.

Author Contributions: Investigation, S.S., S.T.; writing, original draft preparation, S.S.; writing, review and editing, S.K.; supervision, S.T., S.K.

Funding: This research received no external funding.

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

References

1. Turaga, P.; Chellappa, R.; Subrahmanian, V.; Udrea, O. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Vid. Technol.* **2008**, *18*, 1473–1488. [CrossRef]
2. Moeslund, T.; Hilton, A.; Krüger, V. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.* **2006**, *104*, 90–126. [CrossRef]
3. Poppe, R. A survey on vision-based human action recognition. *Image Vis. Comput.* **2010**, *28*, 976–990. [CrossRef]
4. Micucci, D.; Mobilio, M.; Napoletano, P. Unimib shar: A dataset for human activity recognition using acceleration data from smartphones. *Appl. Sci.* **2017**, *7*, 1101. [CrossRef]
5. Yurtman, A.; Barshan, B.; Fidan, B. Activity recognition invariant to wearable sensor unit orientation using differential rotational transformations represented by quaternions. *Sensors* **2018**, *18*, 2725. [CrossRef] [PubMed]
6. Kantoch, E. Recognition of sedentary behavior by machine learning analysis of wearable sensors during activities of daily living for telemedical assessment of cardiovascular risk. *Sensors* **2018**, *18*, 3219. [CrossRef] [PubMed]
7. Chieu, H.; Lee, W.; Kaelbling, L. Activity Recognition from Physiological Data Using Conditional Random Fields. Workshop at ICML. Available online: https://dspace.mit.edu/handle/1721.1/30197 (accessed on 14 November 2018).
8. Zhang, Y.; Peterson, B.; Dong, Z. A support-based reconstruction for sense mri. *Sensors* **2013**, *13*, 4029–4040. [CrossRef] [PubMed]

9. Xu, X.; Tang, J.; Zhang, X.; Liu, X.; Zhang, H.; Qiu, Y. Exploring techniques for vision based human activity recognition: Methods, systems, and evaluation. *Sensors* **2013**, *13*, 1635–1650. [CrossRef] [PubMed]
10. Jalal, A.; Kamal, S.; Kim, D. Shape and motion features approach for activity tracking and recognition from kinect video camera. In Proceedings of the 2015 IEEE 29th International Conference on IEEE, Advanced Information Networking and Applications Workshops (WAINA), Taipei, Taiwan, 27–29 March 2015; pp. 445–450.
11. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Learning actionlet ensemble for 3D human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 914–927. [CrossRef] [PubMed]
12. 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 IEEE, Computer Vision and Pattern Recognition Workshops (CVPRW), San Francisco, CA, USA, 13–18 June 2010; pp. 9–14.
13. Xia, L.; Chen, C.-C.; Aggarwal, J. View invariant human action recognition using histograms of 3d joints. In Proceedings of the 2012 IEEE Computer Society Conference on IEEE, Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, RI, USA, 16–21 June 2012; pp. 20–27.
14. Gaglio, S.; Re, G.; Morana, M. Human activity recognition process using 3-d posture data. *IEEE Trans. Hum.-Mach. Syst.* **2015**, *45*, 586–597. [CrossRef]
15. Aggarwal, J.; Ryoo, M. Human activity analysis: A review. *ACM Comput. Surv. (CSUR)* **2011**, *43*, 16. [CrossRef]
16. Cheng, G.; Wan, Y.; Saudagar, A.; Namuduri, K.; Buckles, B. Advances in human action recognition: A survey. *arXiv* **2015**, arXiv:1501.05964.
17. Aggarwal, J.; Cai, Q. Human motion analysis: A review. *Comput. Vis. Image Underst.* **1999**, *73*, 428–440. [CrossRef]
18. Gavrila, D. The visual analysis of human movement: A survey. *Comput. Vis. Image Underst.* **1999**, *73*, 82–98. [CrossRef]
19. Zhu, F.; Shao, L.; Xie, J.; Fan, Y.G. From handcrafted to learned representations for human action recognition: A survey. *Image Vis. Comput.* **2016**, *55*, 42–52. [CrossRef]
20. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A Survey. *Image Vis. Comput.* **2017**, *60*, 4–21. [CrossRef]
21. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on CVPR 2009 IEEE, Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
22. Yilmaz, A.; Javed, O.; Shah, M. Object tracking: A survey. *Acm Comput. Surv. (CSUR)* **2006**, *38*, 13. [CrossRef]
23. Zhan, B.; Monekosso, D.; Remagnino, P.; Velastin, S.; Xu, L.-Q. Crowd analysis: A survey. *Mach. Vis. Appl.* **2008**, *19*, 345–357. [CrossRef]
24. Weinland, D.; Ronfard, R.; Boyer, E. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.* **2011**, *115*, 224–241. [CrossRef]
25. Aggarwal, J. Motion analysis: Past, present and future. In *Distributed Video Sensor Networks*, Springer: Berlin, Germany, 2011; pp. 27–39.
26. Chaaraoui, A.; Climent-Pérez, P.; Flórez-Revuelta, F. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Syst. Appl.* **2012**, *39*, 10873–10888. [CrossRef]
27. Metaxas, D.; Zhang, S. A review of motion analysis methods for human nonverbal communication computing. *Image Vis. Comput.* **2013**, *31*, 421–433. [CrossRef]
28. Vishwakarma, S.; Agrawal, A. A survey on activity recognition and behavior understanding in video surveillance. *Vis. Comput.* **2013**, *29*, 983–1009. [CrossRef]
29. Cedras, C.; Shah, M. Motion-based recognition a survey. *Image Vis. Comput.* **1995**, *13*, 129–155. [CrossRef]
30. Johansson, G. Visual perception of biological motion and a model for its Analysis. *Percept. Psychophys.* **1973**, *14*, 201–211. [CrossRef]
31. Marr, D.; Nishihara, H. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B* **1978**, *200*, 269–294. [CrossRef] [PubMed]
32. Hogg, D. Model-based vision: A program to see a walking person. *Image Vis. Comput.* **1983**, *1*, 5–20. [CrossRef]
33. Rohr, K. Towards model-based recognition of human movements in image sequences. *CVGIP Image Underst.* **1994**, *59*, 94–115. [CrossRef]

34. Gavrila, D.; Davis, L. Towards 3-d model-based tracking and recognition of human movement: A multi-view approach. In Proceedings of the International workshop on automatic face-and gesture-recognition, Zurich, Switzerland, 26–28 June 1995; pp. 272–277.
35. Green, R.; Guan, L. Quantifying and recognizing human movement patterns from monocular video images-part I: A new framework for modeling human motion. *IEEE Trans. Circuits Syst. Video Technol.* **2004**, *14*, 179–190. [CrossRef]
36. Carlsson, S.; Sullivan, J. Action recognition by shape matching to key frames. In Proceedings of the Workshop on Models Versus Exemplars in Computer Vision, Tokyo, Japan, 18–22 November 2001; Volume 1.
37. Ogale, A.; Karapurkar, A.; Guerra-Filho, G.; Aloimonos, Y. View-Invariant Identification of Pose Sequences for Action Recognition. Available online: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwig37CPp9PeAhXO7GEKHfkGDQ4QFjABegQIBRAC&url=https%3A%2F%2Fpdfs.semanticscholar.org%2F98cb%2F29ae950ee4d3d9f23af0def90c9c3bfc771b.pdf&usg=AOvVaw1ste9TR_jRriyo-ytbTn_V (accessed on 14 November 2018).
38. Rittscher, J.; Blake, A. Classification of human body motion. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 1, pp. 634–639.
39. Darrell, T.; Pentland, A. Space-time gestures. In Proceedings of the 1993 CVPR'93 IEEE Computer Society Conference on IEEE, Computer Vision and PatternRecognition, New York, NY, USA, 15–17 June 1993; pp. 335–340.
40. Yamato, J.; Ohya, J.; Ishii, K. Recognizing human action in time-sequential images using hidden markov model. In Proceedings of the 1992 IEEE CVPR'92 Computer Society Conference on IEEE, Computer Vision and Pattern Recognition, Champaign, IL, USA, 15–18 June 1992; pp. 379–385.
41. Bobick, A.; Davis, J. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 257–267. [CrossRef]
42. Gorelick, L.; Blank, M.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time Shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2247–2253. [CrossRef] [PubMed]
43. Yilmaz, A.; Shah, M. Actions sketch: A novel action representation. In Proceedings of the IEEE CVPR 2005 Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 984–989.
44. Elgammal, A.; Shet, V.; Yacoob, Y.; Davis, L. Learning dynamics for exemplar-based gesture recognition. In Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, WI, USA, 18–20 June 2003; Volume 1, p. I-I.
45. Weinland, D.; Boyer, E. Action recognition using exemplar-based embedding. In Proceedings of the CVPR 2008 IEEE Conference on IEEE Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–7.
46. Lv, F.; Nevatia, R. Single view human action recognition using key pose matching and viterbi path searching. In Proceedings of the CVPR'07 IEEE Conference on IEEE, Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
47. Sminchisescu, C.; Kanaujia, A.; Metaxas, D. Conditional models for contextual human motion recognition. *Comput. Vis. Image Underst.* **2006**, *104*, 210–220. [CrossRef]
48. Zhang, Z.; Hu, Y.; Chan, S.; Chia, L.-T. Motion context: A new representation for human action recognition. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2008; pp. 817–829.
49. Nelson, R.; Polana, R. Qualitative recognition of motion using temporal texture. *CVGIP Image Underst.* **1992**, *56*, 78–89. [CrossRef]
50. Polana, R.; Nelson, R. Low level recognition of human motion (or how to get your man without finding his body parts). In Proceedings of the 1994 IEEE Workshop on IEEE, Motion of Non-Rigid and Articulated Objects, Austin, TX, USA, 11–12 November 1994; pp. 77–82.
51. Cutler, R.; Turk, M. View-based interpretation of real-time optical flow for gesture recognition. In Proceedings of the Third IEEE International Conference on IEEE, Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April1998; pp. 416–421.
52. Efros, A.; Berg, A.; Mori, G.; Malik, J. *Recognizing Action at a Distance Null*; IEEE: Piscataway, NJ, USA, 2003; p. 726.

53. Robertson, N.; Reid, I. Behaviour understanding in video: A combined method. In Proceedings of the ICCV 2005 Tenth IEEE International Conference on Computer Vision, Beijing, China, 17–21 October 2005; Volume 1, pp. 808–815.
54. Wang, Y.; Sabzmeydani, P.; Mori, G. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *Human Motion–Understanding, Modeling, Capture and Animation*; Springer: Berlin, Germany, 2007; pp. 240–254.
55. Zelnik-Manor, L.; Irani, M. Event-based analysis of video. In Proceedings of the CVPR 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; Volume 2, p. II.
56. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the CVPR 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.
57. Thurau, C.; Hlavác, V. Pose primitive based human action recognition in videos or still images. In Proceedings of the IEEE Conference on CVPR Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
58. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. In Proceedings of the CVPR 2008 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2008; pp. 1–8.
59. Laptev, I.; Pérez, P. Retrieving actions in movies. In Proceedings of the ICCV 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
60. Tran, D.; Sorokin, A. Human activity recognition with metric learning. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2008; pp. 548–561.
61. Morguet, P.; Lang, M. Spotting dynamic hand gestures in video image sequences using hidden markov models. In Proceedings of the ICIP 98, 1998 International Conference on IEEE Image Processing, Chicago, IL, USA, 4–7 October 1998; pp. 193–197.
62. Laptev, I. On space-time interest points. *Int. J. Comput. Vis.* **2005**, *64*, 107–123. [CrossRef]
63. Harris, C.; Stephens, M. A combined corner and edge detector. In *Alvey Vision Conference*; Citesee: Manchester, UK, 1988; Volume 15, p. 10-5244.
64. Willems, G.; Tuytelaars, T.; Van Gool, L. An efficient dense and scale-invariant spatio-temporal interest point detector. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2008; pp. 650–663.
65. Dollár, P.; Rabaud, V.; Cottrell, G.; Belongie, S. Behavior recognition via sparse spatio-temporal features. In Proceedings of the 2005 2nd Joint IEEE International Workshop on IEEE, Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, 15–16 October 2005; pp. 65–72.
66. Messing, R.; Pal, C.; Kautz, H. Activity recognition using the velocity histories of tracked keypoints. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 104–111.
67. Matikainen, P.; Hebert, M.; Sukthankar, R. Trajectons: Action recognition through the motion analysis of tracked features. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), Kyoto, Japan, 29 September–2 October 2009; pp. 514–521.
68. Klaser, A.; Marszałek, M.; Schmid, C. A spatio-temporal descriptor based on 3d-gradients. In Proceedings of the BMVC 2008-19th British Machine Vision Conference, British Machine Vision Association, Leeds, UK, 28–29 September 2008; p. 275.
69. Dalal, N.; Triggs, B.; Schmid, C. Human detection using oriented histograms of flow and appearance. In *European Conference on Computer Vision*; Springe: Berlin, Germany, 2006; pp. 428–441.
70. Kantorov, V.; Laptev, I. Efficient feature extraction, encoding and classification for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2593–2600.
71. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.-L. Action recognition by dense trajectories. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 3169–3176.
72. Vig, E.; Dorr, M.; Cox, D. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2012; pp. 84–97.

73. Jiang, Y.-G.; Dai, Q.; Xue, X.; Liu, W.; Ngo, C.-W. Trajectory-based modeling of human actions with motion reference points. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2012; pp. 425–438.
74. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the 2013 IEEE International Conference on IEEE, Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 3551–3558.
75. Peng, X.; Zou, C.; Qiao, Y.; Peng, Q. Action recognition with stacked fisher vectors. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2014; pp. 581–595.
76. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2010; pp. 143–156.
77. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv* **2015**, arXiv:1405.3531.
78. Lan, Z.; Lin, M.; Li, X.; Hauptmann, A.; 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.
79. Sutskever, I.; Vinyals, O.; Le, Q. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2014; pp. 3104–3112.
80. 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 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
81. Li, P.; Wang, D.; Wang, L.; Lu, H. Deep visual tracking: Review and experimental comparison. *Pattern Recognit.* **2018**, *76*, 323–338. [CrossRef]
82. Srinivas, S.; Sarvadevabhatla, R.; Mopuri, K.; Prabhu, N.; Kruthiventi, S.; Babu, R. A taxonomy of deep convolutional neural nets for computer vision. *Front. Robot. AI* **2016**, *2*, 36. [CrossRef]
83. Ciresan, D.; Meier, U.; Gambardella, L.; Schmidhuber, J. Convolutional neural network committees for handwritten character classification. In Proceedings of the 2011 International Conference on IEEE, Document Analysis and Recognition (ICDAR), Beijing, China, 18–21 September 2011; pp. 1135–1139.
84. Cireşan, D.; Meier, U. Multi-column deep neural networks for offline handwritten chinese character classification. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015; pp. 1–6.
85. Kim, S.; Hori, T.; Watanabe, S. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In Proceedings of the 2017 IEEE International Conference on IEEE, Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 4835–4839.
86. Wu, Z.; Valentini-Botinhao, C.; Watts, O.; King, S. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In Proceedings of the 2015 IEEE International Conference on IEEE, Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 4460–4464.
87. Kim, Ho.; Lee, J.S.; Yang, H.-S. Human action recognition using a modified convolutional neural network. In *International Symposium on Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2007.
88. Jones, J.P.; Palmer, L.A. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **1987**, *58*, 1233–1258. [CrossRef] [PubMed]
89. Jhuang, H.; Serre, T.; Wolf, L.; Poggio, T. A biologically inspired system for action recognition. In Proceedings of the IEEE 11th International Conference on ICCV 2007 Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
90. Fukushima, K.; Miyake, S.; Ito, T. Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE Trans. Syst. Man Cybern.* **1983**, *5*, 826–834. [CrossRef]
91. Mutch, J.; Lowe, D.G. Multiclass object recognition with sparse, localized features. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; Volume 1, pp. 11–18.
92. Serre, T.; Wolf, L.; Poggio, T. *Object Recognition with Features Inspired by Visual Cortex*; Massachusetts Inst of Tech Cambridge Dept of Brain and Cognitive Sciences: Cambridge, MA, USA, 2006.
93. 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]

94. Varol, G., Laptev, I.; Schmid, C. Long-term temporal convolutions for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1510–1517. [CrossRef] [PubMed]
95. Ng, J.Y.-H.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
96. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
97. 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, Barcelona, Spain, 7–13 December 2015; pp. 4489–4497.
98. Sun, L.; Jia, K.; Yeung, D.-Y.; Shi, B. Human action recognition using factorized spatio-temporal convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 7–13 December 2015; pp. 4597–4605.
99. Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; Baskurt, A. Sequential deep learning for human action recognition. In *International Workshop on Human Behavior Understanding*; Springer: Berlin, Germany, 2011; pp. 29–39.
100. 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.
101. Robinson, A.; Fallside, F. Static and dynamic error propagation networks with application to speech coding. In *Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1988; pp. 632–641.
102. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2014; pp. 568–576.
103. Brox, T.; Bruhn, A.; Papenberg, N.; Weickert, J. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2004; pp. 25–36.
104. Collobert, R.; Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 160–167.
105. Krizhevsky, A.; Sutskever, I.; Hinton, G. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2012; pp. 1097–1105.
106. 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.
107. Wang, L.; Qiao, Y.; Tang, X. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4305–4314.
108. Sánchez, J.; Perronnin, F.; Mensink, T.; Verbeek, J. Image classification with the fisher vector: Theory and practice. *Int. J. Comput. Vis.* **2013**, *105*, 222–245. [CrossRef]
109. Srivastava, N.; Mansimov, E.; Salakhudinov, R. Unsupervised learning of video representations using lstms. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–1 July 2015; pp. 843–852.
110. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2014; pp. 2672–2680.
111. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
112. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, ACM, Helsinki, Finland, 5–9 July 2008; pp. 1096–1103.
113. Wang, L. Recognition of human activities using continuous autoencoders with wearable sensors. *Sensors* **2016**, *16*, 189. [CrossRef] [PubMed]

114. Liou, C.-Y.; Cheng, W.-C.; Liou, J.-W.; Liou, D.-R. Autoencoder for words. *Neurocomputing* **2014**, *139*, 84–96. [CrossRef]
115. Shin, H.-C.; Orton, M.; Collins, D.; Doran, S.; Leach, M. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 1930–1943. [CrossRef] [PubMed]
116. Yan, X.; Chang, H.; Shan, S.; Chen, X. Modeling video dynamics with deep dynencoder. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2014; pp. 215–230.
117. Mathieu, M.; Couprie, C.; LeCun, Y. Deep multi-scale video prediction beyond mean square error. *arXiv* **2015**, arXiv:1511.05440.
118. Chaquet, J.; Carmona, E.; Fernández-Caballero, A. A survey of video datasets for human action and activity recognition. *Comput. Vis. Image Underst.* **2013**, *117*, 633–659. [CrossRef]
119. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local svm approach. In Proceedings of the 17th ICPR 2004 International Conference on Pattern Recognition, Cambridge, UK, 26–26 August 2004; Volume 3, pp. 32–36.
120. Fisher, R.; Santos-Victor, J.; Crowley, J. Caviar: Context Aware Vision Using Image-Based Active Recognition. 2005. Available online: https://homepages.inf.ed.ac.uk/rbf/CAVIAR/ (accessed on 14 November 2018).
121. Rodriguez, M.; Ahmed, J.; Shah, M. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In Proceedings of the CVPR 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
122. Reddy, K.; Shah, M. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.* **2013**, *24*, 971–981. [CrossRef]
123. Soomro, K.; Zamir, A.; Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2015**, arXiv:1212.0402.
124. Marszalek, M.; Laptev, I.; Schmid, C. Actions in context. In Proceedings of the CVPR 2009 IEEE Conference on IEEE, Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 2929–2936.
125. Nagel, W.; Kröner, D.; Resch, M. *High Performance Computing in Science and Engineering'17*; Springer: Berlin, Germany, 2018.
126. Li, Y.; Li, W.; Mahadevan, V.; Vasconcelos, N. Vlad3: Encoding dynamics of deep features for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1951–1960.
127. Lev, G.; Sadeh, G.; Klein, B.; Wolf, L. Rnn fisher vectors for action recognition and image annotation. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2016; pp. 833–850.
128. 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, 26 June–1 July 2016.
129. Wu, Z.; Jiang, Y.-G.; Wang, X.; Ye, H.; Xue, X.; Wang, J. Fusing multi-stream deep networks for video classification. *arXiv* **2015**, arXiv:1509.06086.