applied sciences

MDPI

Article

Human Behavior Analysis: A Survey on Action Recognition

Bruno Degardin * and Hugo Proença *

IT-Instituto de Telecomunicações, University of Beira Interior, 6201-001 Covilhã, Portugal
* Correspondence: bruno.degardin@ubi.pt (B.D.); hugomcp@di.ubi.pt (H.P.)

Abstract: The visual recognition and understanding of human actions remain an active research domain of computer vision, being the scope of various research works over the last two decades. The problem is challenging due to its many interpersonal variations in appearance and motion dynamics between humans, without forgetting the environmental heterogeneity between different video images. This complexity splits the problem into two major categories: action classification, recognising the action being performed in the scene, and spatiotemporal action localisation, concerning recognising multiple localised human actions present in the scene. Previous surveys mainly focus on the evolution of this field, from handcrafted features to deep learning architectures. However, this survey presents an overview of both categories and respective evolution within each one, the guidelines that should be followed and the current benchmarks employed for performance comparison between the state-of-the-art methods.

Keywords: action detection; biometrics; human action recognition; human activity analysis

Citation: Degardin, B.; Proença, H. Human Behavior Analysis: A Survey on Action Recognition. *Appl. Sci.* **2021**, *11*, 8324. https://doi.org/10.3390/app11188324

Academic Editor: Lidia Jackowska-Strumillo

Received: 3 August 2021
Accepted: 4 September 2021
Published: 8 September 2021

1. Introduction

The advancements in computer technology have allowed the exponential growth of the machine Learning domain. Particularly, the improvements in artificial neural networks enabled diverse research for hard-coded knowledge, whereas the deep learning [1,2] field became the current mainstream. Deep learning solves the main problem of extracting high-level and abstract features, such as every individual pixel when analysing images of persons, where the factors of variations become erratic. Those multiple processing layers dramatically improved the state-of-the-art in solving these problems [3,4], resulting in its increased use in various scientific research domains while bringing breakthroughs in deep convolutional neural networks in processing images, video, speech and audio. The recognition of human actions has become one of the most promising applications of computer vision, due to the continuous advent of image capture equipment and surveillance systems over the last two decades, producing massive video content. In biometrics, in contrast to gait recognition, action recognition should be generalised over small variations within the person's appearance, background clutter, viewpoints and action execution.

The sophistication in behaviour analysis led to the hierarchical arrangement regarding different levels of abstraction, introduced and used by several early reviews in this field [5–7] and also by recent ones [8,9], with the following taxonomy: action primitive, action and activity, ordered in accordance to its complexity. Reporting the atomic movement that can describe the limb level as an action primitive (left leg forward, right arm folding), and describing the whole-body movement as a juncture of action primitives an action (running, jumping). Furthermore, at the highest level of abstraction, composed of several subsequent actions, an activity (jumping hurdles, throwing a football, catching keys from the ground), giving an interpretation of all the movements that are being performed within the image.

The extraction of human dynamics information from image sequences can be further divided into two major image representation categories: local and global representations. In a bottom-up fashion, local representations are based on the detection of spatio-temporal

interest points first and local patches are encoded around these points, combining all the patches into a final representation, alike to the current 3-dimensional convolutional neural networks approaches (*I3D* [10], *C3D* [11] and *R(2+1)D* [12]). Despite being less sensitive to noise and partial occlusion (without requiring background subtraction or tracking of humans), they depend on the extraction of a sufficient amount of relevant interest points, and despite its high accuracy they also lose the global view of the present humans within an image, as they tend to generalise over the several possible different actions being performed. Within the action classification paradigm, most methods applying only local representations will fail when observing multiple actions at the same time. On the other hand, in a top-down fashion, global representations consist of first localising a person in the image and encoding the region of interest (ROI) as the image descriptor, similar to object detectors and tracking methods to localise humans and keep track of its localisation through the image sequences. Despite being powerful representations, they rely on accurate localisation, and consequently, are more sensitive to viewpoint, noise and occlusions. Within the spatiotemporal action localisation paradigm, global representation approaches can discriminate better over different actions performed at the same time over the captured scene. However, those methods are slightly more complex, taking into account its difficulty in distinguishing coexisting human actions.

Early reviews within the area of vision-based human behaviour analysis and recognition, such as Moeslund et al. [5], Turaga et al. [7] and Poppe [6], give a solid overview regarding the a priori deep learning methods over the recognition of human actions and activities, describing the fundamental concepts, techniques and models that were the foundation of the human activity analysis challenges.

After the early stages of the exponential growth of deep learning, Zhu et al. [9] presented one of the first comprehensive surveys which explored the advancements of human behaviour analysis representations, distinguishing the image representations into handcrafted features and learning-based representations (which included deep learning architectures). With the same approach to the action recognition challenge, Herath et al. [8] also discussed a distinction between pioneering handcrafted representations and deep learning techniques, and presented a difference between local representations and global/holistic representations. More recently, Kong et al. [13] presented an extensive and complete survey regarding not only action recognition, but also action prediction, presenting the state-of-the-art evolution on both problems. Table 1 represents an overview of well-known surveys based on topologies, taxonomies and applications.

Table 1. Previous surveys on human behaviour analysis.

				Topologies		Taxonomies		Applications	
		References	Year	Handcrafted	Data-Driven	Actions	Activities	Recogntion	Prediction
Deep Learning	Prior	Moeslund et al. [5]	2006	✓	✗	✓	✓	✓	✗
		Turaga et al. [7]	2008	✓	✗	✓	✗	✓	✗
		Poppe [6]	2010	✓	✗	✓	✗	✓	✗
	Posterior	Zhu et al. [9]	2016	✓	✓	✓	✗	✓	✗
		Herath et al. [8]	2017	✓	✓	✓	✗	✓	✗
		Kong et al. [13]	2018	✓	✓	✓	✗	✓	✓
		This survey	-	✓	✓	✓	✗	✓	✗

The purpose of this work is to provide a comprehensive review of human action recognition by emphasising two major categories (local and global representations), their

evolution on each one and the current state-of-the-art methods employed to achieve a high-level understanding of video image data in each category (Section 2). Additionally, we present the reported results of several must-know methods in Section 4 with corresponding datasets description (Section 3). Some insights about future directions are addressed in Section 5, and finally a conclusion about the topic is given in Section 6.

2. Human Action Recognition

Video data have been in the scope of the computer vision community for decades, resulting in multiple problems such as abnormal event detection [14], person re-identification [15], action recognition [16], video retrieval [17] and many others have been proposed regarding video representations. Human action recognition consists of the extraction of concise features, from video image data, to achieve a high-level understanding allowing computers to recognise human behaviour. Over the last decade, significant improvements were accomplished through the emerging deep learning models, distinguishing two categorizations in terms of feature descriptors, local and global representations.

2.1. Local Representations

As previously discussed, local representations are composed of a collection of local descriptors, which are sampled from space-time interest points, as observed in Figure 1.

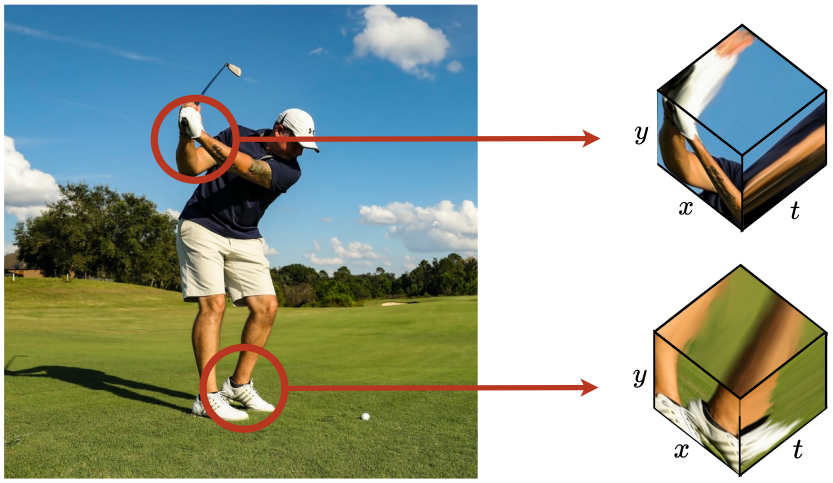
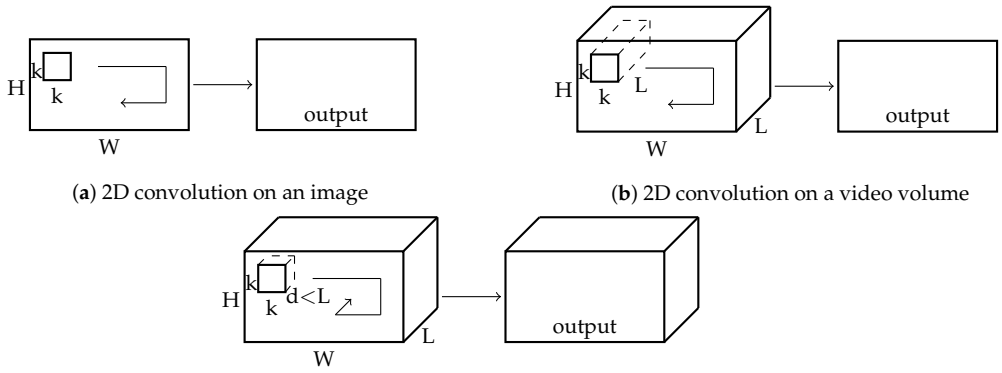


Figure 1. Extraction of local space-time cuboids at interest points in image sequences.

Inspired by the deep learning breakthroughs in the image domain, it is proposed by Tran et al. [11] a spatio-temporal feature learning by using deep convolutional 3-dimensional networks (3D ConvNets). Justified by its better extraction to model temporal information [11,18–20] in comparison to the conventional deep 2-dimensional convolutional networks (2D ConvNets), Tran et al. [11] also employed a deconvolution method [21] to understand and visualise what C3D was learning internally. The difference between those convolutional operations are illustrated in Figure 2, where the application of 2D convolution over an image and over multiple images (video image data) will output an image. Therefore, using 2D ConvNets, most of the networks lose their input's temporal signal after every convolution operation. On the other hand, 3D convolution will better preserve the temporal information, as it does not operate only spatially, but also temporarily, obtaining an output volume as a result. 2D and 3D pooling operations employ the same phenomena.



(**a**) 2D convolution on an image (**b**) 2D convolution on a video volume

(**c**) 3D convolution

Figure 2. 2D and 3D convolution operations (adapted from the work in [11]).

As 3D ConvNets are being increasingly used for the extraction of human dynamics, several variants have been introduced [22–26]. With the application of 3-dimensional convolutional networks, recent approaches had a general focus on combining multiple features, apart from only images. Exploiting the use of optical flow, Carreira and Zisserman [10] employed the Inception-v1 architecture [27] (CNN architecture with multiple size filters operating at the same level) with ImageNet [28] as their backbone network. They improved the 3D ConvNets performance by including an optical-flow stream. Moreover, also using the Inception module [27], but this time only with RGB information, Wang et al. [25] applied an LSTM network [29] analysing the output features from the Inception 3D ConvNet (I3D) to better model the temporal information. Due to the importance of the holistic view in action recognition, Diba et al. [30] applied 3D ConvNets to extract temporal information and merged a second stream of 2D ConvNets, also in order to extract its spatial structure in the frame.

Despite the 3D ConvNets performance, there are still some competitive approaches extracting the spatial and temporal information separately. Zhu et al. [31] proposed an end-to-end trainable two-stage approach, where one stream is responsible for estimating the well-known and powerful technique of optical flow, projecting its motion information to a second network and analysing its temporal information to predict the action label. Then, with a second stream, they extract the spatial information also to predict the action label and applying a late fusion over the weighted average of the predictions scores from both streams. Moreover, also achieving similar performance to 3D ConvNets, Lin et al. [32] on top of a 2D ConvNet, proposed a temporal shift module (TSM), which shifts some parts of the temporal channels in order to exchange information among adjacent frames (shifting one-quarter of the channels due to the low performance and efficiency of a full shifting). They introduced the unidirectional (online) shift that exchange temporal information from the previous frames to the future frames, and also the bidirectional (offline) shift where the mixing is applied in both past frames and future frames. Using ResNet-50 [33] as their backbone network, they apply the temporal shift, from T frames, inside the residual block and before the convolution operation, not affecting the spatial feature learning capability as the activation information is the original.

Recently, motivated by 2D ConvNets, which remain solid performers in action recognition, Tran et al. [12] factorised the 3D convolutional filters into separate spatial and temporal components. This spatiotemporal decomposition, shown in Figure 3, splits the computation into a spatial 2D convolution with a temporal 1D convolution afterwards. In a simplified manner, this new convolution can be interpreted as the analysis of the temporal information from t frames sequence with a kernel size of 1, after a conventional 2D convolution from one image. Moreover, also in the kernel factorisation paradigm, Xie et al. [34] factorised, in some convolution filters, the Inception module [27] similar to the $(2+1)$D block in Figure 3. This spatiotemporal kernel factorisation improved the performance significantly over regular 3D ConvNets and inspired further developments

on $(2+1)$D convolutions. Likewise, using the ResNet-50 [33] as the backbone network, Qiu et al. [35] proposed three architecture variants, denominated as P3D, applying the $(2+1)$D convolution inside the residual blocks: The first one in a cascade manner, similar to Figure 3, where the two kinds of filters influence each other over the same path. A second architecture, where the spatial and temporal filters are operated in a parallel fashion, being directly accumulated at the end. Additionally, a third design is proposed, where the spatial 2D filters are directly accumulated to the output of the block, and the spatial filters influence the temporal 1D filters being also accumulated to the output. Despite the first proposal achieving higher accuracy, they also presented a complete version, mixing all the three variants, achieving even higher accuracy. Furthermore, Qiu et al. [35] applied DeepDraw [36], inside the P3D ResNet model, to visualise the class knowledge of some categories.

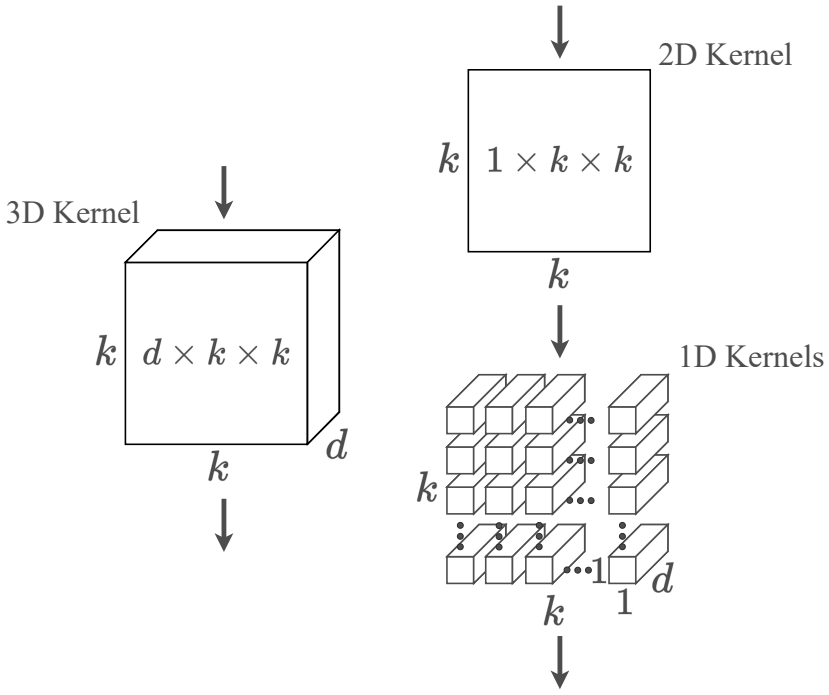


Figure 3. 3D vs (2 + 1)D. Factorisation of the spatiotemporal 3D convolution into a spatial 2D convolution followed by temporal 1D convolutions.

Under the local representation fashion, Qiu et al. [24] operated over local and global diffusion (LGD) blocks, defined as a local and global path of feature extraction interacting with each other, to capture better large-range dependencies. Their local path exploits the P3D [35] as the local transformation, and the global path is obtained from a global average pooling (GAP) of the local feature. Then, in the subsequent local layer, the global feature is upsampled to formulate the global priority. Consequently, they are not only able to classify the action in frame-wise manner, but also in a pixel-wise one by taking into account the global view of the video clip, extracting the ROIs from the local feature, and performing spatiotemporal action localisation.

The temporal global average pooling (TGAP) layer used at the end of almost all 3D CNNs [11,12], extract the final temporal information's richness. However, the prior features from TGAP represent the different temporal regions of a clip, where some parts of the temporal feature might be more important and beneficial than others, and taking its simple averaging may not be the best choice. Therefore, Kalfaoglu et al. [37] proposed an attention mechanism denoted as bidirectional encoder representations from transformers (BERT) [38], which provided unprecedented success on natural language processing (NLP), here applied for better temporal modelling. Composed of a positional encoding from the temporal

features to preserve positional information, and applying a position-wise feedforward network to learn a better subspace for the attention mechanism and classification. Their BERT attention employing R(2+1)D [12] architecture is the current state-of-the-art in local representations for action recognition (Section 4).

2.2. Global Representations

Taking into account the holistic view of the scene, which may include different actions simultaneously, the image representation is described as a global representation. By capturing the motion information of the entire human subject, global representations are richer and express better and more concise motion information. Although they are susceptible to noise, the current advances in human detector [39–42], human tracker [43–46], and multi-person tracker [47–50] algorithms, make it easier to achieve high accuracy even with occlusions, different viewpoints, or noise, as shown in Figure 4. Despite the object detectors and trackers accuracy, they capture the information in a certain rectangle region, which may introduce some noise and irrelevant information, not only from the human appearance but also the cluttered background. Therefore, in order to take advantage of those powerful algorithms, usually, some earlier feature extraction is required, rather than using a raw input of person's localisation for the extraction of human dynamics.

Figure 4. Multi-person tracking example employing the FAIRMOT tracker [50].

Following a region proposal network (RPN), Peng et al. [51] proposed a spatial RPN analysing one frame and a motion RPN analysing the optical flow of its neighbouring frames (flow of 5 frames). Their architecture was based on faster R-CNN [52] for region proposals, and all the regions from both streams are fused before the ROI pooling layer. Resorting to the single-shot multibox detector (SSD) framework [40], Kalogeiton et al. [53] extend the anchor boxes to anchor cuboids over subsequent frames, extracting the 2D convolutional features with shared weights between frames. Engaging 3D ConvNets, Gu et al. [54] extract motion information through the analysis of two-streams, RGB frames and optical flow of the clip with an Inception 3D ConvNet. They employed faster R-CNN [52] for region proposals, applying ROI pooling on both branches of their network, and average pooling is used at the feature map level to fuse them. Recognising human dynamics as a regression problem, Köpüklü et al. [55] employed a 3D ResNext-101 [56] to extract temporal information from a clip video and use a 2D-CNN branch on the most recent frame of the clip to address the spatial localisation, stacking both resulting features from the networks and following the same guidelines as YOLOv2 [57] for the bounding box regression. Employing a progressive learning framework, Yang et al. [58], in order to refine the cuboid proposals towards spatiotemporal action localisation, proposed a multi-step optimisation process to refine initial proposals progressively. They used a two-stream architecture for spatial refinement and temporal extent, where the spatial branch performs bounding box regression at each frame, taking into account the temporal extent in order to update the proposals regarding the cuboids extension through a 3D ConvNet. Moreover, also analysing cuboids, Li et al. [59] presented an action tubelet (cuboid) detector, denoted as a moving centre detector. Treating an action tubelet instance as a trajectory of moving points, they employed a three-branch framework, where the centre branch detects the action instance centre and classification. The movement branch estimates the offset estimation in the current frame concerning its centre, and finally, the box branch predicts the bounding box size over the predicted centre point. Feichtenhofer et al. [60] exploited both spatial and temporal information through different frame rates over a two-stream architecture. A fast and a slow pathway, where the fast one (high frame rate) will extract

temporal information through a 3D ConvNet, and the slow one (at low frame rate) will analyse only spatial information taking into account the temporal dynamics. Its slow pathway is able to localise an action based on the fast pathway.

Nevertheless, commonly using object detectors and trackers as its foundation, one of the most promising human representations is the extraction of multi-person pose estimation [61–66], as shown in Figure 5. Human skeleton sequences have three distinguishing characteristics: Starting with the existence of strong correlations between each node and adjacent nodes, consequently, skeleton frames are rich in body structural information. Second, its temporal continuity exists across frames within the same joints and also in the body structure, and, last but not least, a co-occurrence relationship between spatial and temporal domains is present in that kind of data. Furthermore, this technique overcomes all appearance noises that human region proposals can contain, being modular, semantically rich and very descriptive, and consequently, driving the learning process of the model exclusively on human behaviour.

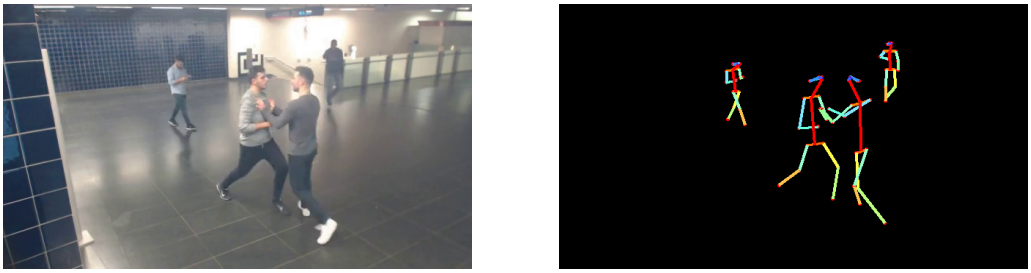


Figure 5. Multi-person pose estimation example using DensePose [63] from the Detectron2 framework [64].

One of the earliest methods that explored skeleton data for action recognition was the work by Junejo et al. [67], where they explored the self-similarity matrix (SSM), which is computed by the distances between action representations of all pairs of time frames. They claimed that the SSMs are approximately invariant under viewpoint changes, as illustrated in Figure 6. Applying different types of features to compute the SSM, they concluded that between the same feature type, the pattern similarity was effectively similar.

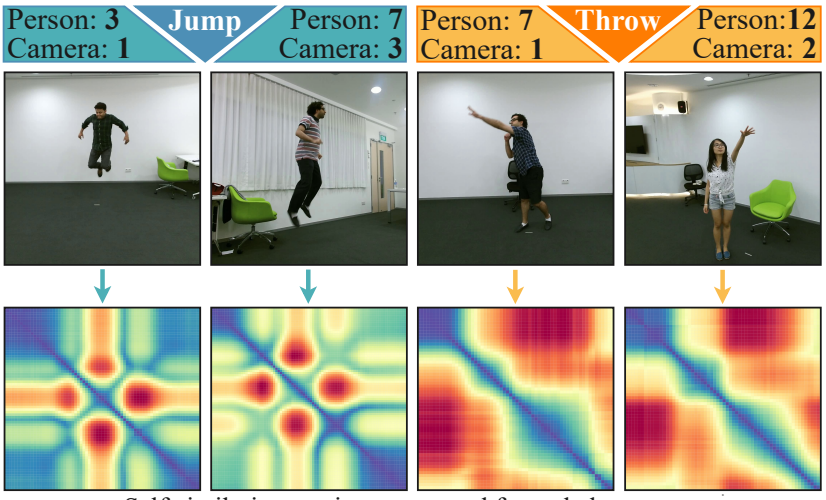


Figure 6. The first row illustrates two actions performed by different subjects and under different poses (jump up in the first two columns and throw in the two rightmost columns). The bottom row provides the corresponding SSM yielding from the corresponding skeleton sequences (adapted from the work in [68]).

As a structured data type, some methods employed LSTM networks [29] to model the time-series. Exploring this algorithm, Liu et al. [69] proposed to convert the pose estimated to a tree structure in order to be unfolded as a sequence. Then, each LSTM unit is fed with a skeletal joint, which also takes into account the neighbouring joints and previous frames of the same joint. When analysing the human pose performing some actions in the real world, usually some skeleton joints have more importance than others, paying different attention to different regions of the scene [70]. Song et al. [71], in the same field of LSTM networks, proposed to model skeleton joints in a selective way as an attention mechanism. Composed of two attention networks, the spatial one assumes the weight of a joint (its importance) as the resulting activations from the network, and the temporal attention one uses the input gate of the LSTM network for learning to control the amount of information (its importance) to be used, in each frame, for the final classification decision. In a similar way, Zhang et al. [72] also proposed a recurrent neural network [73] with LSTM, but this time they take into account the translation from global body movement (the whole body dynamics in the scene) to local body posture (skeleton configuration upon the body centre in the first frame). This way, it is possible to adapt its viewpoint in order to be a more suitable observation for orientation alignment normalisation.

Even though skeleton pose estimation is a structured data type, several methods approached the problem with 2D ConvNets [74–77]. Li et al. [77] proposed a two-stream 2D ConvNet: one to extract features from spatial coordinates of the pose in a 3D manner (position, joints and frames) through a skeleton transformer module, which extracts weighted interpolated joints matrix. On the other stream, they extract the skeleton motion through computed distances between frames. Ke et al. [75] presented a new representation for skeleton data, employing cylindrical coordinates generating a collection of clips which are used as input to a CNN.

More recently, as an emerging topic in deep learning research, generalising neural networks towards structured graph data resulted in graph convolutional networks (GCNs) [78–82]. Justified by its better extraction of concise features among graph structured data, GCNs have been in the scope of several works towards action recognition with skeleton data [68,83–86]. Usually, a spatiotemporal graph convolution is defined as a set of nodes and edges, where the nodes represent the skeleton joints and the edges denote the connectivity between those joints intra-frame and inter-frame. Figure 7 illustrates a GCN architecture example using skeleton data.

One of the first methods to develop a spatiotemporal GCN, for human behaviour understanding, was the work by Yan et al. [86] where they presented three partition strategies (neighboring). Uni-labelling gives the same vector weight to all neighbour joints; however, they can lose the local differential over the skeleton sequence. Distance Partitioning yields two weight vectors for the root node and the remaining neighbours, extracting local differential properties. At last, spatial configuration partitioning, which labels the nodes according to their distance from the gravity centre of the skeleton. Instead of using undirected graphs, where the GCN will learn its connections by itself, Shi et al. [87] proposed a directed GCN in order to model the dependencies of joints and bones in the human body to extract local information better. Despite the effectiveness of considering the skeleton joints dependencies, there must have flexibility, in order to the network extract its own relevant dependencies from the skeleton features. Si et al. [88] proposed an LSTM aggregated to a GCN with the purpose of better extracting its temporal information, which they use for the selection of key joints in order to produce a soft attention mechanism. Tang et al. [85] proposed a reinforcement learning [89] strategy combined with a GCN for action recognition. Their agent is responsible for extracting the most informative frames (keyframes) in order to feed the GCN more efficiently. With the applicability of the recent technique of neural architecture search (NAS) [90], Peng et al. [91] proposed a dynamic GCN, where its connectivity is built upon a search space based on node correlations, achieving competitive results with its state-of-the-art approaches (Section 4).

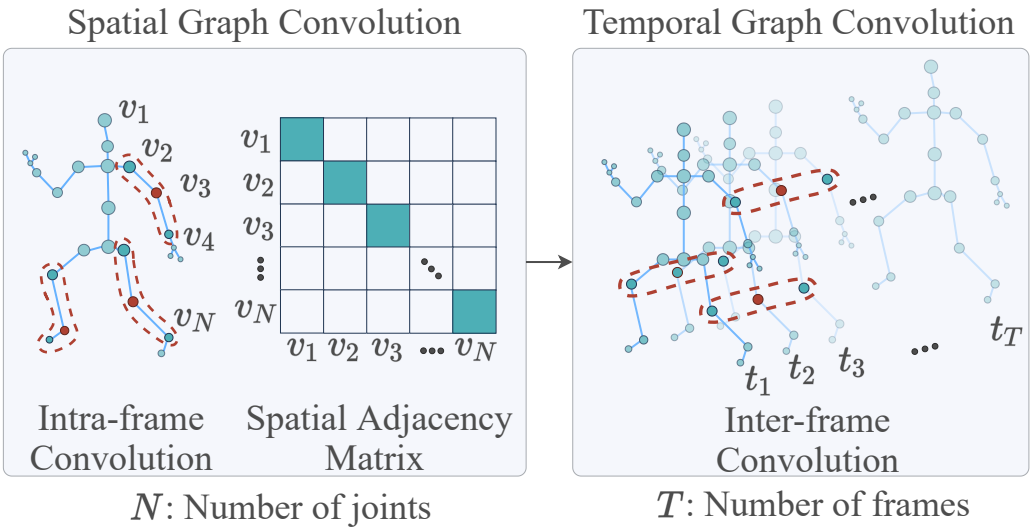


Figure 7. A spatiotemporal graph of a skeleton sequence. Light green dots represent the body joints (graph nodes). Light blue edges illustrate the intra-body edges. The spatial graph convolution receives as input a skeleton graph with its corresponding adjacency matrix to control the intra-frame (spatial) convolution (red dotted line) from the root node (red joint) neighbourhood. Then, 1-dimensional convolution is performed on the same positional joints across consecutive frames, resulting in the temporal (inter-frame) convolution.

Considering the evolution of deep convolutional neural networks, Hinton et al. [92] introduced capsule networks as a new representation method that successfully overcame the state-of-the-art in some problems. A set of neurons composes a capsule, where its activity vector represents different features of a specific type of entity. A capsule network follows a level hierarchy, where higher-level capsules will cover more extensive regions of the image (more complex representations with more degrees of freedom) while in the counterpart, the lower-level capsules will make predictions for smaller regions of the image, with the rationale that when multiple low-level capsules achieve a prediction consensus, a higher-level capsule will become active. Inspired by the advances in capsule networks [93], Duarte et al. [94] proposed a capsule network analysing the 3-dimensional data in order to achieve spatiotemporal action localisation. Following a masking procedure, the capsule activations are set to 0, except for the capsule representing the ground truth class, predicting the action localisation through the largest activation and feeding a fully-connected network in order to extract a feature map for better localisation.

3. Datasets

The current benchmarks present a wide diversity of different controlled sequences, environments and feature extraction exploration. This section will present the most popular ones among their respective categories, where most of the popular state-of-the-art methods (reported here) are competing. We divided the datasets corresponding to their respective evaluation protocol, such as frame-level (mostly employed by local representation approaches) and pixel-level (mostly employed by global representation approaches).

3.1. Frame-Level Benchmarks

UCF-101 [95] consists of 13,320 realistic videos widely collected from Youtube, containing 101 action classes with a wide diversity in intra-class and inter-class, and large variations of camera motion, object scale, object appearance, cluttered backgrounds, viewpoints, different illuminations. This dataset provides the frame-level ground-truth of the actions from all videos and is one of the most popular benchmarks among the action recognition methods at the frame-level.

HMDB-51 [96] with 51 action categories, is composed of 7000 clips from Youtube videos to digitised movies, where each class contains at least 101 videos, providing a great diversity between action classes. This dataset provides the frame-level ground-truth of the actions from all videos and is also one of the most popular datasets for evaluation at the frame-level.

Kinetics-400 [97] has 400 human action classes, where each action has at least 400 video samples from Youtube, and each video clip has a duration of 10 seconds. With great heterogeneity, this dataset provides the frame-level ground-truth of all videos' actions. Being a well-known dataset, the authors released two extended versions, the Kinetics-600 [98] with 600 action categories, where each action has at least 600 clips, and very recently, the Kinetics-700 [99] with 700 action classes, where each class has at least 700 videos. However, the 400 version is currently the most popular one of these three versions.

The **Sports-1M** dataset [19] is currently the largest video dataset composed of 1,133,158 videos, which have been annotated automatically with 487 action categories at the video-level, presenting an extreme diversity of sports videos. However, its availability is only provided through individual video URLs, making it difficult to access the videos.

THUMOS'14 [100] consists of approximately 18,000 videos widely collected containing 101 action classes, providing its ground-truth labels at the frame-level. This dataset has the peculiarity of providing only trimmed videos for the training phase, and methods should be evaluated on untrimmed data over the validation and test set.

3.2. Pixel-level Benchmarks

UCF-101-24 [95] is the second version of ground-truth labels from the original UCF-101, where they provide the bounding box annotations of the humans present in the videos. Although there are 101 classes, these pixel-level labels only represent 24 classes of them. This dataset is one of the most popular benchmarks among action recognition methods at the pixel-level.

J-HMDB-21 [101] is the second version of ground-truth annotations from the original HMDB-51, where they provide the bounding box labels of the humans present in the videos. These labels at the pixel-level represent 21 action categories from the original 51. This benchmark is also one of the most popular datasets for evaluation at the pixel-level.

AVA [54] (atomic visual actions) is composed of 430 video clips (15 minutes each) from different movies, containing 80 atomic visual actions. Following the same activity hierarchy as previously mentioned (Section 1), the ground-truth labels (provided at the pixel-level) of this dataset represent the atomic body movements or object manipulations at its lowest possible level of natural descriptions, such as the pose action (sit, stand, run, etc.), object interaction (if applicable, carry, write, ride etc.), and person-to-person (if applicable, talk to, listen to, watch, etc.).

NTU RGB+D [102] consists of 56,880 video samples with 60 action classes. This dataset was captured from highly restricted camera views providing 3D skeleton and RGB-D data for each video sample. This benchmark was built for the purpose of exploring the skeleton dynamics of the human body, not only for its estimation but also to recognise the action performed, being one of the most popular for evaluation of skeleton-based action recognition methods. A second version of this dataset was recently released, NTU RGB+D 120 [103], adding 60 classes and 57,600 video clips to the original version.

Kinetics-Skeleton [97] was introduced by skeleton-based action recognition methods, ignited by the challenging diversity of the Kinetics-400 dataset, action recognition methods based on skeleton data started employing multi-person pose estimators [61–63,65] in order to extract its skeleton data to feed their models.

4. Evaluation Protocols and Quantitative Analysis for Action Recognition

In this section, we provide a performance comparison in Table 2 over a comprehensive list of 18 must-know methods in each category addressed in this survey, which each method was explained in Sections 2.1 and 2.2. The results are reported on six challenging

benchmarks, being the most popular datasets for evaluation comparison among each category. Likewise, the performance measures reported are the most typical ones for each category approach. The accuracies are directly reported from the original works.

The evaluation protocol for local representation approaches is frame-level recognition, reporting the Top-1 accuracy as the performance measure (the average accuracy regarding the Top-1 class predicted by the model). For the global representation approaches, the evaluation protocol performed for action recognition is pixel-wise, adopting as the performance measure the mean average precision (mAP), which approximates the area under the precision–recall curve for each individual action class. Additionally, we also indicate the year of the method regarding when it was published.

Aside from the intra-representations performance evolution, we can observe a significant difference between performances of local and global representations regarding RGB-based datasets (UCF-101 [95], HMDB-51 [96], UCF-101 24 [95] and J-HMDB-21 [101]). This is justified by the difficulty of the problem being solved. As described in Section 1 local representations are performing action recognition at the frame level, while global representations are performing at the pixel level, which becomes far more challenging. Despite skeleton-based methods working at the pixel level and achieving great performance on NTU RGB-D [102], this dataset was obtained from highly restricted settings. When applied to a more wild and challenging dataset, such as Kinetics-Skeleton [97], a notable drop in performance is observed, which indicates an important limitation of skeleton-based methods.

Table 2. Performance summary of some reference action recognition methods from both categories, local and global representation approaches over their respective benchmarks in terms of accuracy and mean average precision.

	Method	**Year**	**Top-1**	**Top-1**
			UCF-101	*HMDB-51*
Local Representations	Slow Fusion [19]	2014	0.654	-
	C3D [11]	2015	0.823 †	-
	TS-LSTM [104]	2019	0.943	0.690
	H Two-stream I3D [31]	2018	0.971	0.787
	R(2+1)D [12]	2018	0.973 †	0.787 †
	HTNet [30]	2019	0.978	0.765
	Two-stream I3D [10]	2017	0.979	0.802
	R(2+1)D - BERT [37]	2020	0.987	0.851
	Method	**Year**	**mAP$_{50}$**	**mAP$_{50}$**
			UCF-101 24	*J-HMDB-21*
Global Representations	STEP [58]	2019	0.750	-
	Faster R-CNN + I3D [54]	2018	0.763	0.733
	MOC [59]	2020	0.780 †	0.708 †
	VideoCapsuleNet [94]	2018	0.786 †	0.646 †
	YOWO [55]	2019	0.804 †	0.757 †
			NTU RGB-D	*Kinetics-S*
	ST-LSTM [69]	2016	0.755	-
	ST-GCN [86]	2018	0.883 †	0.307 †
	GCN-NAS [91]	2020	0.957	0.371
	DGNN [87]	2019	0.961	0.369

† Reproduced applying the original's author configuration to confirm the results reported. "-" Result is not reported.

5. Current Challenges, Trends, and Further Directions

Human understanding through video image data has been exponentially improved since temporal information extraction through the emerging of 3-dimensional convolutional networks. However, most of the current approaches employ multiple branches, analysing different features to produce richer and more robust information. On the other hand, some methods employ backbone networks for the initial feature extraction (temporal or regional), dividing both training and inference process into a two-stage process each. Despite its high effectiveness, the inference time its sacrificed, and most of the methods do not even achieve a ten frame rate. This problem is relatively more serious for global representation approaches, as they tend to predict multiple actions simultaneously. Therefore, some future breakthroughs are required in order to develop unified architectures for action recognition, which will significantly reduce the inference time, increase its speed, and make it easier for embedded devices.

As previously discussed, the current benchmarks are very extensive, such as the Sports-1M dataset, the AVA dataset, among others. Consequently, the video annotation process becomes an extremely exhausting task concerning the unpredictable number of video hours needed to successfully train a model. Therefore, there is a need for semi-supervised and unsupervised learning algorithms towards the recognition of human actions. The problem resides in the high complexity of this family of algorithms, and without forgetting, the increasing number of action classes becomes even more challenging due to the higher overlapping between classes. This problem could be tackled by recognising simple basic actions at first, such as walking, running, and jumping, not achieving a high-level of human behaviour understanding as existing supervised methods, but it could be a starting point to be improved in the future.

The human's surrounding contextualization is regarded as the Achilles' heel in understanding human behaviour. Considering the presence of objects in the scene (alongside or being manipulated by humans), the extraction of spatial information concerning the background clutter, and the interpretability of human interactions between multiple humans. There is a lack of focus in this direction as the complexity of the problem increases, and current approaches are still improving the individual action recognition. However, as a future direction, once a method achieves reasonable performance, those contextualizations could be encoded through knowledge-based approaches or statistical models, such as finite-state automatons and Markov models, where nodes or states would contain information about the observed human behaviour and verification of detected objects or background identification. Moreover, they could also be encoded through syntactic approaches, such as grammars and dictionary algorithms, where activities (junction of subsequent actions) are treated in a cascade manner. Therefore, achieving the highest level of abstraction, as previously mentioned, identifying activities.

6. Conclusions

Over the last decade, deep learning had an evident impact on the improvements towards action recognition. However, several conceptual breakthroughs would be needed in order to achieve another exponential growth and overcome the current limitations. In this paper, we provided an overview concerning human behaviour analysis, presenting state-of-the-art techniques and must-know methods in this field. The explained concepts and methods were divided into local and global representations to clarify their distinction in solving similar challenges. Over the last years, those image representation approaches were merged to extract even more concise features from video image data and achieve a higher level of understanding from the observed scene's behaviour.

Despite the maturity of visual recognition and perception of human actions, effective deployment of this kind of technology in fully unconstrained scenarios is still far away.

Author Contributions: B.D.'s contributions are in conceptualisation, methodology, data curation, formal analysis, validation and writing—original draft preparation. H.P. performed project administration, conceptualisation, funding acquisition and manuscript revision. Both authors have read and agreed to the published versions of the manuscript.

Funding: This work is funded by FCT/MEC through national funds and co-funded by FEDER-PT2020 partnership agreement under the project UIDB//50008/2020. Also, it was supported by operation Centro-01-0145-FEDER-000019-C4-Centro de Competências em Cloud Computing, co-funded by the European Regional Development Fund (ERDF) through the Programa Operacional Regional do Centro (Centro 2020), in the scope of the Sistema de Apoio à Investigação Científica e Tecnológica-Programas Integrados de IC&DT, and supported by 'FCT-Fundação para a Ciência e Tecnologia' through the research grant 'UI/BD/150765/2020'.

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

References

1. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
2. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.
3. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105.
4. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117.
5. Moeslund, T.B.; Hilton, A.; Krüger, V. A survey of advances in vision-based human motion capture and analysis. *Comput. Vis. Image Underst.* **2006**, *104*, 90–126.
6. Poppe, R. A survey on vision-based human action recognition. *Image Vis. Comput.* **2010**, *28*, 976–990.
7. Turaga, P.; Chellappa, R.; Subrahmanian, V.S.; Udrea, O. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1473–1488.
8. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **2017**, *60*, 4–21.
9. Zhu, F.; Shao, L.; Xie, J.; Fang, Y. From handcrafted to learned representations for human action recognition: A survey. *Image Vis. Comput.* **2016**, *55*, 42–52.
10. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6299–6308.
11. 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.
12. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.
13. Kong, Y.; Fu, Y. Human action recognition and prediction: A survey. *arXiv* **2018**, arXiv:1806.11230.
14. Degardin, B.; Proença, H. Iterative weak/self-supervised classification framework for abnormal events detection. *Pattern Recognit. Lett.* **2021**, *145*, 50–57.
15. Wang, G.; Lai, J.; Huang, P.; Xie, X. Spatial-temporal person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8933–8940.
16. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 21–23 May 2013; pp. 3551–3558.
17. Bendersky, M.; Garcia-Pueyo, L.; Harmsen, J.; Josifovski, V.; Lepikhin, D. Up next: Retrieval methods for large scale related video suggestion. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014, pp. 1769–1778.
18. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231.
19. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
20. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 568–576.
21. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 818–833.
22. Duan, H.; Zhao, Y.; Xiong, Y.; Liu, W.; Lin, D. Omni-sourced Webly-supervised Learning for Video Recognition. *arXiv* **2020**, arXiv:2003.13042.

23. Hong, J.; Cho, B.; Hong, Y.W.; Byun, H. Contextual action cues from camera sensor for multi-stream action recognition. *Sensors* **2019**, *19*, 1382.
24. Qiu, Z.; Yao, T.; Ngo, C.W.; Tian, X.; Mei, T. Learning spatio-temporal representation with local and global diffusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12056–12065.
25. Wang, X.; Miao, Z.; Zhang, R.; Hao, S. I3d-lstm: A new model for human action recognition. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, UK, 2019; Volume 569, p. 032035.
26. Yang, H.; Yuan, C.; Li, B.; Du, Y.; Xing, J.; Hu, W.; Maybank, S.J. Asymmetric 3d convolutional neural networks for action recognition. *Pattern Recognit.* **2019**, *85*, 1–12.
27. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.
28. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
29. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
30. Diba, A.; Fayyaz, M.; Sharma, V.; Paluri, M.; Gall, J.; Stiefelhagen, R.; Van Gool, L. Holistic large scale video understanding. *arXiv* **2019**, arXiv:1904.11451.
31. Zhu, Y.; Lan, Z.; Newsam, S.; Hauptmann, A. Hidden two-stream convolutional networks for action recognition. In *Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 363–378.
32. Lin, J.; Gan, C.; Han, S. Tsm: Temporal shift module for efficient video understanding. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 7083–7093.
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
34. Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 305–321.
35. Qiu, Z.; Yao, T.; Mei, T. Learning spatio-temporal representation with pseudo-3d residual networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 Ocotber 2017; pp. 5533–5541.
36. DeepDraw. Available online: https://github.com/auduno/deepdraw (accessed on 5 September 2021).
37. Kalfaoglu, M.; Kalkan, S.; Alatan, A.A. Late temporal modeling in 3d cnn architectures with bert for action recognition. *arXiv* **2020**, arXiv:2008.01232.
38. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
39. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
40. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
41. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
42. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
43. Chen, B.X.; Tsotsos, J.K. Fast visual object tracking with rotated bounding boxes. *arXiv* **2019**, arXiv:1907.03892.
44. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 19–20 June 2019; pp. 4282–4291.
45. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High performance visual tracking with siamese region proposal network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8971–8980.
46. Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; Torr, P.H. Fast online object tracking and segmentation: A unifying approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 19–20 June 2019; pp. 1328–1338.
47. Bergmann, P.; Meinhardt, T.; Leal-Taixe, L. Tracking without bells and whistles. In Proceedings of the IEEE International Conference on Computer Vision, Long Beach, CA, USA, 19–20 June 2019; pp. 941–951.
48. Brasó, G.; Leal-Taixé, L. Learning a neural solver for multiple object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6247–6257.
49. Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; Wang, S. Towards real-time multi-object tracking. *arXiv* **2019**, arXiv:1909.12605.
50. Zhan, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. A Simple Baseline for Multi-Object Tracking. *arXiv* **2020**, arXiv:2004.01888.
51. Peng, X.; Schmid, C. Multi-region two-stream R-CNN for action detection. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 744–759.

52. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99.
53. Kalogeiton, V.; Weinzaepfel, P.; Ferrari, V.; Schmid, C. Action tubelet detector for spatio-temporal action localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 Ocotber 2017; pp. 4405–4413.
54. 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, 18–23 June 2018; pp. 6047–6056.
55. Köpüklü, O.; Wei, X.; Rigoll, G. You Only Watch Once: A Unified CNN Architecture for Real-Time Spatiotemporal Action Localization. *arXiv* **2019**, arXiv:1911.06644.
56. Hara, K.; Kataoka, H.; Satoh, Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6546–6555.
57. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 Ocotber 2017; pp. 7263–7271.
58. Yang, X.; Yang, X.; Liu, M.Y.; Xiao, F.; Davis, L.S.; Kautz, J. Step: Spatio-temporal progressive learning for video action detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 19–20 June 2019; pp. 264–272.
59. Li, Y.; Wang, Z.; Wang, L.; Wu, G. Actions as Moving Points. *arXiv* **2020**, arXiv:2001.04608.
60. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE International Conference on Computer Vision, Long Beach, CA, USA, 19–20 June 2019; pp. 6202–6211.
61. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 Ocotber 2017; pp. 7291–7299.
62. Neverova, N.; Novotny, D.; Vedaldi, A. Correlated Uncertainty for Learning Dense Correspondences from Noisy Labels. 2019. Available online: https://openreview.net/forum?id=SklKNNBx8B (accessed on 5 September 2021).
63. Riza Alp Güler, Natalia Neverova, I.K. DensePose: Dense Human Pose Estimation in the Wild. 2018. Available online: https://openaccess.thecvf.com/content_cvpr_2018/html/Guler_DensePose_Dense_Human_CVPR_2018_paper.html (accessed on 5 September 2021).
64. Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.Y.; Girshick, R. Detectron2. 2019. Available online: https://github.com/facebookresearch/detectron2 (accessed on 5 September 2021).
65. Xiu, Y.; Li, J.; Wang, H.; Fang, Y.; Lu, C. Pose Flow: Efficient Online Pose Tracking. arXiv **2018**, arXiv:1802.00977.
66. Zhang, Z. Microsoft kinect sensor and its effect. *IEEE Multimed.* **2012**, *19*, 4–10.
67. Junejo, I.N.; Dexter, E.; Laptev, I.; PÚrez, P. Cross-view action recognition from temporal self-similarities. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 293–306.
68. Degardin, B.; Lopes, V.; Proença, H. REGINA-Reasoning Graph Convolutional Networks in Human Action Recognition. *arXiv* **2021** , arXiv:2105.06711.
69. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 816–833.
70. Goferman, S.; Zelnik-Manor, L.; Tal, A. Context-aware saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1915–1926.
71. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. *arXiv* **2016** , arXiv:1611.06067.
72. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 Ocotber 2017; pp. 2117–2126.
73. Graves, A. Supervised sequence labelling. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 5–13.
74. Jia, J.G.; Zhou, Y.F.; Hao, X.W.; Li, F.; Desrosiers, C.; Zhang, C.M. Two-Stream Temporal Convolutional Networks for Skeleton-Based Human Action Recognition. *J. Comput. Sci. Technol.* **2020**, *35*, 538–550.
75. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. A new representation of skeleton sequences for 3d action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 Ocotber 2017; pp. 3288–3297.
76. Kim, T.S.; Reiter, A. Interpretable 3d human action analysis with temporal convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Venice, Italy, 22–29 Ocotber 2017; pp. 1623–1631.
77. Li, C.; Zhong, Q.; Xie, D.; Pu, S. Skeleton-based action recognition with convolutional neural networks. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017; pp. 597–600.
78. Duvenaud, D.K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R.P. Convolutional networks on graphs for learning molecular fingerprints. *arXiv* **2015**, arXiv:1509.09292.
79. Henaff, M.; Bruna, J.; LeCun, Y. Deep convolutional networks on graph-structured data. *arXiv* **2015**, arXiv:1506.05163.
80. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.

81. Li, Y.; Tarlow, D.; Brockschmidt, M.; Zemel, R. Gated graph sequence neural networks. *arXiv* **2015**, arXiv:1511.05493.
82. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S.Y. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4–24.
83. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 19–20 June 2019; pp. 3595–3603.
84. Si, C.; Jing, Y.; Wang, W.; Wang, L.; Tan, T. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 103–118.
85. Tang, Y.; Tian, Y.; Lu, J.; Li, P.; Zhou, J. Deep progressive reinforcement learning for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Munich, Germany, 8–14 September 2018; pp. 5323–5332.
86. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv* **2018**, arXiv:1801.07455.
87. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with directed graph neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 19–20 June 2019; pp. 7912–7921.
88. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 19–20 June 2019; pp. 1227–1236.
89. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
90. Zoph, B.; Le, Q.V. Neural architecture search with reinforcement learning. *arXiv* **2016**, arXiv:1611.01578.
91. Peng, W.; Hong, X.; Chen, H.; Zhao, G. Learning Graph Convolutional Network for Skeleton-Based Human Action Recognition by Neural Searching. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 2669–2676.
92. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3856–3866.
93. Hinton, G.E.; Sabour, S.; Frosst, N. Matrix capsules with EM routing. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
94. Duarte, K.; Rawat, Y.; Shah, M. Videocapsulenet: A simplified network for action detection. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 7610–7619.
95. 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.
96. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
97. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
98. Carreira, J.; Noland, E.; Banki-Horvath, A.; Hillier, C.; Zisserman, A. A short note about kinetics-600. *arXiv* **2018**, arXiv:1808.01340.
99. Smaira, L.; Carreira, J.; Noland, E.; Clancy, E.; Wu, A.; Zisserman, A. A Short Note on the Kinetics-700-2020 Human Action Dataset. *arXiv* **2020**, arXiv:2010.10864.
100. Jiang, Y.G.; Liu, J.; Roshan Zamir, A.; Toderici, G.; Laptev, I.; Shah, M.; Sukthankar, R. THUMOS Challenge: Action Recognition with a Large Number of Classes. 2014. Available online: http://crcv.ucf.edu/THUMOS14/ (accessed on 5 September 2021).
101. Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; Black, M.J. Towards understanding action recognition. In Proceedings of the IEEE international conference on computer vision, Sydney, Australia, 1–8 December 2013; pp. 3192–3199.
102. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1010–1019.
103. Liu, J.; Shahroudy, A.; Perez, M.L.; Wang, G.; Duan, L.Y.; Chichung, A.K. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2684–2701.
104. Ma, C.Y.; Chen, M.H.; Kira, Z.; AlRegib, G. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Process. Image Commun.* **2019**, *71*, 76–87.