sensors

MDPI

Review

Intelligent Video Analytics for Human Action Recognition: The State of Knowledge

Marek Kulbacki [1,2,*], Jakub Segen [1,2], Zenon Chaczko [2,3], Jerzy W. Rozenblit [4], Michał Kulbacki [2], Ryszard Klempous [5] and Konrad Wojciechowski [1]

[1] Polish-Japanese Academy of Information Technology, 02-008 Warsaw, Poland
[2] DIVE IN AI, 53-307 Wroclaw, Poland
[3] School of Electrical and Data Engineering, University of Technology Sydney, Ultimo 2007, Australia
[4] Department of Electrical and Computer Engineering, The University of Arizona, Tucson, AZ 85721, USA
[5] Wrocław University of Science and Technology, 50-370 Wroclaw, Poland
* Correspondence: mk@pja.edu.pl

Abstract: The paper presents a comprehensive overview of intelligent video analytics and human action recognition methods. The article provides an overview of the current state of knowledge in the field of human activity recognition, including various techniques such as pose-based, tracking-based, spatio-temporal, and deep learning-based approaches, including visual transformers. We also discuss the challenges and limitations of these techniques and the potential of modern edge AI architectures to enable real-time human action recognition in resource-constrained environments.

Keywords: intelligent video analytics; edge AI; visual transformers; human activity recognition; video surveillance; pose-based HAR; tracking-based HAR; spatio-temporal-based HAR; deep learning-based HAR

check for updates

Citation: Kulbacki, M.; Segen, J.; Chaczko, Z.; Rozenblit, J.W.; Kulbacki, M.; Klempous, R.; Wojciechowski, K. Intelligent Video Analytics for Human Action Recognition: The State of Knowledge. *Sensors* **2023**, *23*, 4258. https://doi.org/10.3390/s23094258

Academic Editor: Wataru Sato

Received: 10 December 2022
Revised: 9 April 2023
Accepted: 21 April 2023
Published: 25 April 2023

1. Introduction

Independent security systems, known as security or CCTV cameras, register video footage, and video surveillance cameras monitor specific areas. A single security camera typically produces fifteen to sixty pictures a second, resulting in 3 million images daily. Global information provider IHS Markit reports that in 2015 in the UK alone, 5 million CCTV cameras recorded 25 billion hours of video sequences. Supervisors registered around 350 million operating security cameras worldwide until 2016. In 2017, there were 176 million surveillance cameras in China's private and public sectors, increasing to 2.7 billion by the end of 2022 [1].

Over the past twenty years, video surveillance systems based on CCTV have become a widely used and effective method of deterring, preventing, and detecting crimes [2–4]. Monitoring solutions have migrated from single-unit solutions to intelligent multi-camera network structures, including edge-based architectures of wireless video sensor networks [5] with security and bandwidth constraints [6]. The scope of application and operation of video surveillance systems is extensive, even if limited to human activity recognition. Systems often require complex implementation procedures, resulting in the need to employ specialized surveillance companies to ensure correct security during mass events [7]. Statistical data confirm that the introduction of CCTV systems to monitor areas with an increased level of risk allows for a significant reduction (even by 50%) in the number of robberies and acts of antisocial behaviour by increasing the effectiveness of the services responsible for the implementation of security tasks [8–11]. Nevertheless, the same reports indicate that the current surveillance systems also possess many imperfections: low credibility of the recorded content, often resulting from systems' poor technical capabilities or difficult registration conditions, the unreliability of the human factor, and the immaturity of legislative procedures in monitoring and response. The essence of monitoring concerning the security aspect of citizens should be an efficient flow of complete information between system operators and the police [12,13].

The number of video surveillance system installations and the amount of information collected is rapidly increasing, creating problems in collecting and processing information by supervisors. Research shows that, on average, after twenty minutes of observing one screen, the operator may overlook 90% of what is happening in the monitored place [8]. The current development directions of IP CCTV solutions [5] are systems of intelligent analysis of dynamic scenes with the automatic detection of many moving objects [14] and understanding their behaviours [15–17]. Due to the functional requirements, the market distinguishes the development of active and passive video surveillance systems. Passive systems usually record the monitored zone's video stream for evidence in the event of a crime. Active systems support the supervisors with additional information from the presented or processed image. The most advanced research concerns IVA [18]. Such systems attempt to obtain a description of events in the monitored zone and then take appropriate actions based on the interpretation of monitored events [16]. The necessary image registration and processing are associated with difficulties analogous to those occurring in computer vision systems, remarkably the variability of illumination, observation point, orientation, and distance from the observed object. It is challenging to build a general-purpose intelligent surveillance system [19] and continuously supply it with electrical power [20]. That is why professionals adapt existing systems to specific requirements [21,22]. The most significant difficulty is developing generalized algorithms to solve specific IVA-related problems. Therefore, intelligent surveillance systems usually comprise a library of modules with algorithms for specific applications [23]. The increasingly popular no-code/low-code computer vision platforms reduce the entry threshold into computer vision-based solutions for non-professionals, where applications are built visually from developed components. Gartner predicts [24] that by 2024 more than 65% of applications will be developed with no/low code development.

This survey provides a comprehensive overview of HAR methods chosen specifically for potential use with surveillance cameras, categorizing them into four distinct groups. It also discusses the advantages and disadvantages of each group, including their efficiency and suitability for IVA applications. Finally, it synthesizes the most recent and relevant research on these methods, providing readers with up-to-date insights into the strengths and limitations of each class of methods. The paper offers information on suitable datasets to make models more useful for practical use in intelligent video surveillance challenges. This will make it easier for readers to comprehend the value of data in creating HAR models and allow them to choose relevant datasets for their unique IVA applications. It is a helpful resource for academics and industry professionals who want to enhance the efficiency and dependability of IVA systems for the challenge of human action recognition techniques from the monocular camera in video surveillance.

2. Intelligent Video Analytics

The typical workflow in intelligent video surveillance systems includes the following stages: image acquisition, object and motion detection, object classification, object tracking, analysis and understanding of behaviour and activity, people identification, and information fusion in multi-camera systems [25–31].

Virtually every intelligent surveillance system detects objects and motion. Motion detection requires segmenting adjacent areas representing moving and stationary objects [32]. The most popular approaches to motion segmentation include temporal differencing, background subtraction, and optical flow [25,26,31]. Statistical background subtraction is a more frequently used segmentation method due to its resistance to disturbances, shadows, and changes in lighting [33,34]. Researchers usually use optical flow [35] to detect the movement of a point or specific feature [36] in a video sequence using traditional or modern methods such as FlowFormer [37] or PanoFlow [38]. However, most optical flow approaches are unsuitable for real-time applications because of their vulnerability to interference and difficulty in putting algorithms into practice. The unequivocal detection of moving areas in a video sequence allows attention to focus on these areas during subsequent processes, such as tracking or behaviour analysis, and speed up the entire processing process [39,40]. The subsequent processing steps, including object tracking, behaviour analysis, and recognition, strongly depend on the detection effect.

An unambiguous classification of moving objects is necessary to track them and analyse their behaviour accurately. Classification is understood here as a standard pattern recognition task. The most popular categories of approaches used to classify objects include classification based on recognized shape and motion [26]. Motion classification is sometimes greatly facilitated because, in general, human movement exhibits periodic properties.

Table 1 provides a broader context for using HAR methods in IVA systems by outlining the different workflow elements and operations involved in implementing such solutions. As one can see, the HAR particular classes of recognition strategies described in the paper are just one part of the overall workflow, which includes data acquisition, pre-processing, object detection, object tracking, event detection, decision making, and alert generation. Depending on the specific use case, some or all of these elements may be present in an IVA system. Tracking an object in a system with many cameras in real-time under changing conditions is a complicated task [28]. The object tracking task uses the classification results. It becomes more complex when more moving objects [15,41–43] appear on the scene, which is treated as a background when tracking the selected object [44]. We can treat the tracking problem as a correspondence problem finding a visual object in two consecutive image frames [45]. The position of an object during tracking is usually transformed into 3D coordinates. We can divide tracking methods into four main categories based on [26,44,45]: regions, contours [46], features [47], model, and a hybrid that combines the advantages of region- and feature-based methods or a combination of these methods. Sequential Monte Carlo methods [48], particularly condensation algorithms [49], dominate the group of generalized tracking methods.

Understanding and interpreting movement plays an essential role in intelligent surveillance systems. Recognition of human movement from the video stream starts the process of extracting information about movement from an image sequence. The surveillance system can learn patterns of movement, e.g., walking, extracting the features that determine movement dynamics by decomposition of a tracked movement [50–53]. There are three leading methods of extracting motion information from an image sequence: information from optical flow feature analysis [54], information from trajectory-based features [44,55–57], and information from region-based features [58].

The CCD, thermal imaging, and night vision cameras are the three most popular image recording devices in surveillance systems [19]. Simultaneous acquisition and presentation of images from cameras of various technologies, such as vision and thermal imaging, ensure optimal day/night vision in various weather conditions [59]. The separate processing of image information results in individual and different results with the inherent flaws of each image acquisition technology. In the case of vision cameras, these provide low-contrast images at night, in bad weather, and at long distances. In the case of thermal imaging cameras, low resolution, poor contrast in rainfall as well as ambiguity in the intuitive interpretation of the images from long distances. Data fusion, i.e., the superimposition of images from cameras of different technologies and the presentation of the resulting image on one screen, improves image quality, eliminates the weaknesses of the combined technologies, and increases the efficiency and comfort of the operator [59,60]. In surveillance systems, there is also a need to simultaneously present the image from many cameras partially covering the viewing areas or automatically switch such cameras when the tracked object appears in the field of view of another camera. Using photogrammetric measurement methods from multiple cameras at the initial image processing stage facilitates the exact data fusion process [61,62].

Table 1. Typical workflow elements and operations for implementing IVA systems with HAR methods.

Workflow Element	Operation	HAR Method
Data Acquisition	Capture video data using sensors (e.g., cameras)	-
Pre-processing	Filter, stabilize, and/or enhance the video data	-
Feature Extraction	Extract relevant features from the video data	Pose-Based, Tracking-Based, Spatio-temporal-Based, Deep Learning-Based
Object Detection	Detect objects of interest in the video data	-
Object Tracking	Track objects of interest over time	Tracking-Based
Human Action Recognition	Recognize human actions from video data	Pose-Based, Tracking-Based, Spatio-temporal-Based, Deep Learning-Based
Event Detection	Detect events of interest (e.g., abnormal behaviour)	-
Decision making	Analyse the output of the IVA system and make decisions based on predefined rules	-
Alert Generation	Generate alerts based on the decisions made by the system	-

The assessment of motion detection performance, object tracking [33,63], object classification, intent detection, behaviour, and identification in intelligent video surveillance systems are complex, but the performance impacts the product. The performance is one of the leading topics of the annual challenges around PETS [22], or currently ActivityNet [64] and MMAct [65], with many algorithms, strategies, and benchmark datasets. The 2D PETS datasets include indoor and outdoor human and vehicle tracking with single and multi-camera, posture classification, facial expression, and interaction. ActivityNet is a large-scale activity recognition challenge that aims to cover many complex human activities in people's daily lives. MMAct is a multimodal dataset for action understanding based on diverse modalities.

Companies' current area of interest concerns developing methods and analyses that effectively detect human and object behaviour based on activity patterns [66,67]. Most current video surveillance systems use solutions that allow for effective image processing, mainly in motion or object detection and object tracking in single-camera systems. Object tracking works well in open terrain, but its effectiveness drops when more than one object is in the scene or occlusions occur [68]. The manufacturers of current systems focus mainly on defined patterns of behaviour. This approach shows low effectiveness and often leads to many ambiguities in the events identified by the system. Essential strategies for developing intelligent video surveillance components in the next decade will include interpreting tracked objects in 3D space and advanced real-time behaviour analysis by the adaptive discovery of behaviour patterns.

In recent years, edge processing [69] has grown in popularity, and many large companies have developed small chips to suit the image processing workload. The most well-known products are Google Coral™, Intel Movidius™, NVIDIA Jetson™, Qualcomm Snapdragon™, Apple A-series™, Xilinx Alveo™, and Kneron™. A separate group of deep learning solutions is lightweight image recognition algorithms and the related edge AI trend. The principle of operation of each of the edge processing products is similar. The software includes a dedicated optimizer model that takes models pre-trained in the MXNet [70], TensorFlow [71], Caffe [72], ONNX or other less popular frameworks. The available models can recognize images, human faces, bodies, or objects, on the edge device integrated with the computer system. The software transforms them into a simplified internal representation of a specific architecture. These are hardware architectures for the aforementioned four products: tensor processing unit, vision processing unit, graphics processing unit, and neural processing unit. The particular inference engine loads the simplified representation of the model. The efficiency of inference is so low that it is not yet suitable for solving human action recognition problems in real-time from a video stream of at least at 25/30 fps. However, one of the disadvantages of edge AI recording and analysis is the cost of cameras with sufficient computing power. At the same time, having cheaper cameras and leaving the processing to the server can be less costly.

Many surveillance architectures evolved from the cloud to the edge model and are called hybrids. Edge cloud design [73] is increasingly seen as a natural advancement of cloud computing and an enabler of the coming industrial revolution with the widespread IoT. It includes fog computing, cloudlet, mobile edge, micro data centres, and many others [74]. Video analytics over the cloud offers the benefits of server systems, such as centralized, top-down control, and advanced AI analytics, but without server costs and maintenance needs. It is usually provided as video surveillance as a service (VSaaS) model. Hence, there is no upfront cost, including video recording, storage, remote viewing, management alerts, and cybersecurity. The main differences between the most popular IVA architectures are present in Table 2.

Table 2. Main differences between cloud, hybrid, and edge IVA architectures.

Feature	**IVA Architectures**		
	Cloud	**Hybrid**	**Edge**
System	Centralized	**Decentralized**	**Decentralized**
Power Consumption	**High Compute Power**	**High Compute Power**	Limited Compute Power
Latency	High Latency	**Lowered Latency**	**Lowest Latency**
Bandwidth	High Bandwidth	**Lowered Bandwidth**	**Lowest Bandwidth**
Security	**Secure and Public**	**Secure and Private**	Public
Scalability	**Endless Scalability**	Limited Scalability	Lowest Scalability
Model	**Endless Scalability**	Limited Scalability	Lowest Scalability

The rest of the paper describes different HAR methods that can be used as an algorithmic component of an IVA system. The methods are grouped into four categories: pose-based methods, tracking-based methods, spatio-temporal-based methods, and deep learning-based methods. For each group of methods, the paper discusses the pros and cons of using them for HAR, including their accuracy, efficiency, and suitability for different applications. Overall, the paper provides a comprehensive overview of different HAR methods that can be used in an IVA system and provides valuable insights into each method's strengths and weaknesses, helping readers make informed decisions about which method may be best suited for their particular application.

3. Human Action Recognition Methods

The topic of HAR in videos is an increasingly popular field, as evidenced by the number of publications each month. Google Scholar collated 10,000 scientific articles published between 2020 and 2022. on HAR. More detailed information on HAR provides extensive literature with various methods and comprehensive reviews [75–83]. The following summary presents only the selected aspects and main directions in this area to determine the selection of research directions needed to develop the methods and algorithms necessary to choose the method more effectively.

Workflows of various approaches for HAR can differ significantly, especially regarding the methods used for feature extraction, action segmentation, and classification (Table 3).

Table 3. Typical workflow elements and operations for implementing different HAR.

Workflow Element	Pose-Based Methods	Tracking-Based Methods	Spatio-temporal-Based Methods	Deep Learning-Based Methods
Data Collection	Use sensor devices (e.g., cameras) to capture pose images/videos of humans performing various actions.	Use sensors to capture the movement of the object/person over time.	Use sensors to capture the movement of the object/person over time while also capturing spatial information.	Collect large-scale, labelled datasets of images or videos of humans performing various actions.
Feature Extraction	Extract features from the pose images/videos, such as joint positions, angles, and velocities.	Extract features related to the motion, such as velocity, acceleration, and trajectory.	Extract motion and spatial information features, such as optical flow and histogram of oriented gradients (HOG).	Use deep convolutional neural networks (CNNs) to automatically extract features from raw images or videos.
Action Segmentation	Segment the pose images/videos into individual action sequences.	Segment the object/person movement into individual action sequences.	Segment the object/person movement and spatial information into individual action sequences.	Use recurrent neural networks (RNNs) to segment the video into individual action sequences.
Classification	Classify the individual action sequences into pre-defined action categories.	Classify the individual action sequences into pre-defined action categories.	Classify the individual action sequences into pre-defined action categories.	Use CNNs or RNNs to classify the individual action sequences into pre-defined action categories.
Post-processing	Perform smoothing, filtering, or temporal alignment on the predicted action sequences.	Perform smoothing, filtering, or temporal alignment on the predicted action sequences.	Perform smoothing, filtering, or temporal alignment on the predicted action sequences.	Perform post-processing on the predicted action categories, such as non-maximum suppression or ensembling.

Pose-based methods typically extract features related to joint positions, angles, and velocities and may involve techniques such as skeletonization and joint detection. Action segmentation may include detecting the start and end of action sequences based on changes in common positions or velocities.

Tracking-based methods may use motion-based features such as velocity, acceleration, and trajectory and may involve techniques such as optical flow or motion history images. Action segmentation may involve detecting changes in motion patterns or transitions between different motion patterns.

Spatio-temporal-based methods typically involve extracting features that combine motion and spatial information, such as optical flow or histogram of oriented gradients (HOG). Action segmentation may involve detecting changes in movement and spatial patterns or transitions between different spatial and motion patterns.

Deep learning-based methods typically involve using deep neural networks to learn features from raw image or video data automatically. They may include techniques such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), or transformers. Action segmentation may involve using RNNs to learn temporal patterns in the data and segment action sequences. Thus, while there may be some overlap in the workflows of different HAR approaches, the specific methods and techniques used can differ significantly, depending on the system.

Transfer learning [84] and continual learning [85] are essential concepts in deep learning-based methods for HAR. Transfer learning involves using pre-trained models to transfer knowledge to new tasks. In contrast, continual learning, also known as lifelong learning [86], consists of updating a model incrementally as new data becomes available. Other related concepts include federated learning [87], which involves training a model across decentralized devices, and gossip learning [88] federated variant that does not require an aggregation server; multi-task learning, which consists of preparing a single model to perform multiple associated tasks; ensemble learning [89], which combines various models to make predictions; and reinforcement learning, which involves an agent learning to make decisions in an environment through trial-and-error interactions. Researchers are exploring ways to incorporate these concepts into HAR algorithms to improve performance and efficiency.

The field of visual analysis of human actions and behaviour [90,91] currently has broad practical applications in industry, medicine, and surveillance. The biggest market includes IVA [68,92–94], but also monitoring and supporting systems for ambient-assisted living [95–98] and rapidly growing applications involving visually controlled interactions between people and robots [99–101].

The dynamic development of the evidence management market [102] has increased cameras to over 1.5 million worldwide. According to IDC's predictions [103], the global amount of data created, captured, and replicated worldwide will increase to 175 zettabytes (1 zettabyte = 1 trillion gigabytes) by 2025. There are still many open problems where the task of human action recognition is far from being solved.

3.1. Pose-Based Methods

Detecting, associating, and tracking human skeleton keypoints is a computer vision problem involving human pose estimation [104] and tracking. Significant processing resources required to execute skeleton keypoints tracking in live video data limit the precision of human posture estimation results in real-time. Thanks to recent advancements, new real-time applications are now conceivable. As a result, state-of-the-art approaches often rely on customizing CNN architecture for human posture inference. As depicted in Table 4, the workflow of common elements and operations for human action recognition using pose-based methods consists of several stages, including pose estimation, feature extraction, and classification.

Table 4. Workflow of common elements and operations for human action recognition using pose-based methods.

Workflow Elements	Variety of Operations
1. Pose Estimation	(a) Model-based methods, (b) Deep learning methods, (c) Hybrid methods.
2. Feature Extraction:	(a) Handcrafted features, (b) Deep learning-based features.
3. Classification:	(a) SVM, (b) Random Forest, (c) Neural Networks.

The two most popular strategies for pose-based methods are top-down and bottom-up. The top-down approach begins with a person detector and estimates body parts inside the identified bounding boxes. The bottom-up approach begins by estimating each body part individually, then grouping them to create a unique configuration. Among models suitable for pose-based methods, three are the most popular [105]: Kinematic or skeleton-based for 2D and 3D pose estimation, volumetric for 3D pose estimation, and planar or contour-based composed of one shape or geometric body parts. Recent research has also addressed reliable tracking and pose estimation in natural scenes. Table 5 shows a comparison of several pose estimation methods based on their accuracy, the number of joints they can estimate, their approach (e.g., top-down or bottom-up), and their backbone architecture. The table provides a useful overview of the strengths and weaknesses of each method. The most established 2D real-time multi-person keypoint detection is OpenPose [106], and its faster commercial competitor wrnchAI. Next, are the AlphaPose framework [107,108] and Mask R-CNN [109] based on feature pyramid network (FPN) [110] and a ResNet101 backbone [111]. HRNet [112] maintains high-resolution representation for pose estimation, while DeepCut [113] follows a bottom-up approach, detects people, and subsequently estimates their body configurations. DeepPose captures all joints and uses deep neural network regressors for pose estimation [114], and DensePose maps all human pixels from RGB image to its 3D body surface [115].

Table 5. Summary of selected pose estimation methods on various benchmarks.

Method	Accuracy (%)	Joints	Approach	Backbone
OpenPose [106]	93.8	25	Top-Down and Bottom-up	VGG-19
AlphaPose [107]	87.7	18	Top-Down	ResNet
Mask R-CNN [109]	91.4	17	Top-Down	ResNet
HRNet [112]	95.0	17	Bottom-Up	HRNet
DeepCut [113]	91.0	15	Bottom-Up	VGG
DeepPose [114]	70	16	Top-down	ResNet
DensePose [115]	74.7	24	Top-down	ResNet
MediaPipe [116,117]	88.8	33	Bottom-Up	MobileNet
Yolo [118]	-	17	Bottom-Up	CSPDarknet
Kinect SDK [119]	83.5	25	Top-Down	-
wrnchAI [120]	88.4	57	Bottom-Up	-
PoseNet [121]	86.8	17	Top-Down	MobileNet
ST-GCNs [122]	93.2	18	Bottom-Up	-
AGC-LSTM [123]	94.5	25	Top-down	GC-LSTM

BeomJun et al. [124] compared and analysed the major pose estimation frameworks. Pose-based methods for HAR use an explicit skeletal representation for motion description. The topology of the human skeleton is an important parameter. YOLOv7 [118], a one-shot multi-person pose detector, has a topology with 17 landmarks for a single person, while MediaPipe [116] has 32 keypoints for a single-person skeleton. These methods estimate human pose by identifying skeleton anatomical joints or keypoints in each video frame. Video frames have a sequential nature, so using RNNs, such as Bayesian CG-LSTM [125], hierarchical bi-RNN [126] or AGC-LSTM [123] and graph convolutional networks [122], has made these architectures very common. We do not recommend using skeleton-based representations to describe human-like objects moving in a real-life environment with constraints causing regular silhouette occlusions. We mentioned important works because pose-based methods represent one of the main promising directions for HAR. More information and many open problems connected with this direction are addressed in the following surveys [105,127–131].

3.2. Tracking-Based Methods

While developing an effective system for tracking humans in the video stream, one should address some considerations. The tracking-based methods [132] must be capable of following the tracked human object even under visually difficult situations such as changing illuminations, occlusions, cluttered backgrounds, and complicated human movements, all of which can cause tracking issues. In addition to changes in the environment where a human object is found, the human object can change itself. Such change calls for a consistent tracking system to possess a mechanism that can adapt to the actual human object's appearance. Table 6 provides an overview of the workflow, common elements, and operations used for human action recognition with tracking-based methods.

Table 6. Workflow of common elements and operations for human action recognition using tracking-based methods.

Workflow Elements	Variety of Operations
1. Object Detection	(a) Background subtraction, (b) Haar cascades, (c) Deep learning-based methods.
2. Object Tracking:	(a) Optical Flow, (b) Kalman Filter, (c) Deep learning-based methods.
3. Classification:	(a) SVM, (b) Random forest, (c) Neural Networks.

A system dealing with live video must be able to handle data quickly. The speed of the viewed object determines the processing speed, but at least 25 fps must be established to provide a near-real-time effect. As a result, a quick and efficient implementation is essential, as is the selection of high-performance algorithms. Tracking algorithms can use visual features [133] such as histogram of gradient (HOG) [134] colour [135], Haar [136] and popular learning methods such as support vector machine (SVM) [137], or ensemble learning methods, e.g., boosting [138]. To localize objects, deterministic methods [135] and stochastic methods [138] have been used. Compensation for appearance changes can be achieved using robust mixture models [139] or online boosting [136]. An additional problem is the minimization of the occlusion of surrounding objects [140]. Current reviews of classic object-tracking methods are included in [141,142]. Most of the selected object-tracking methods refer to two categories:

1. Tracking methods using detection (e.g., [143]). For the assignment methods, optimal allocation methods, such as the Hungarian, optimal flow, or graph-based discrete optimization methods, are used. These methods recognize each tracked object in each frame and then group objects from consecutive frames so that each group creates a separate trajectory. The group of tracking methods using detection include:
   (a) a multiphase cascade method with a moving time window [144];
   (b) methods based on a generalized solution of optimal cliques in generalized minimum clique problem (GMCP) graphs [145] and globally optimal generalized maximum multi clique (GMMCP) problems [146];
   (c) a method of generalized linear allocation of short GLA tracklets [147];
   (d) methods based on the estimation of the similarity measure of ADMM dynamics [148] i IHTLS [149];
   (e) a simultaneous tracking method with object segmentation, using a multi-label conditional random field to determine the optimal set of trajectories [150];
   (f) a method to detect and track homogeneous objects with high density [151] using gradients and contours;
   (g) a machine learning method with appearance discrimination using high-certainty tracklets [152], and incremental linear discriminant analysis [153].

2. Methods of tracking using correlations whose main advantage is the speed obtained by using FFT:
    (a) a robust tracking method with accurate scale estimation [154] using a HOG descriptor [134]. The method uses a discriminative correlation filter in the MOSSE method [155]. The method [154] is faster than competitive methods: 2.5 times faster than [156], 25 times faster than ALSA [157], and 250 times faster than SCM [158].
    (b) The method [159] uses dense space-time context learning for tracking.
    (c) The fast-tracking method with kernel correlation filters [160] based on HOG descriptors.

An alternative tracking approach is represented by methods based on clustering trajectories:

1. the method [161] clusters paths constructed from points detected by the SURF detector using the SIFT/SURF descriptor for comparison;
2. the method [42] clusters paths constructed from points of extreme contour curvature.

A typical single- and multi-object tracking approach uses a detector for object localization and re-identification for object association. Hundreds of methods have been competing in SOT [162], GOT [163], MOT, and MOTS challenges since 2015 [164–168]. Recent trends indicate interesting directions into trackers derived from deep learning-based transformers [169–171] applying visual attention [172] for object tracking. Some authors address similar methods for long-term tracking scenarios [173–175]. For the comprehensive, up-to-date summary, this [176] work investigates the present state of DL-based visual tracking algorithms, evaluation metrics, and benchmarks in-depth with leading visual tracking methods.

3.3. Spatio-Temporal Methods

This section provides an overview of the selected and most representative space-time methods for action recognition in temporally and spatially trimmed videos. Algorithms and methods have been improving over the years, and the problem shifted from recognizing actions in videos recorded in laboratories to realistic datasets such as HMDB51 [177], UCF101 [178], Hollywood2 [179] or VMASS [180], MCAD [181], surveillance camera fight [182], and RWF-2000 [183] datasets created directly from surveillance cameras (Section 3.5).

Table 7 shows that deep learning methods can be considered spatio-temporal-based approaches too. There are certainly many other deep learning methods beyond those described in the above table that could be used to analyse human movement. However, the focus of the table is to provide an overview of some commonly used and effective deep learning methods for HAR rather than to provide an exhaustive list. After conducting broad research with many different spatio-temporal-based HAR methods [184], we focus in this section on the bag of visual words (BoVW) approach to presentation among the most promising and actual classic spatio-temporal methods. The BoVW approach has been widely used and benchmarked for human action recognition and is still relevant today. By describing BoVW in detail in the spatio-temporal-based section of the table, we highlight the fact that it is one of the foundational approaches in this field and has been used as a benchmark to compare the performance of newer deep learning methods. For BoVW in a video stream, a part of an image is the visual equivalent of a word, and it can be represented by a bag of quantized invariant local descriptors [185]. This approach provides a flexible choice of processing algorithms using features computed independently over each set of automatically detected RoIs. Finally, such a method structure is easily scalable and robust to the occlusion of motion regions, representing people or partial visibility in a video stream. Figure 1 presents the BoVW approach pipeline divided into ten steps. First, the videos are acquired and annotated for a supervised learning problem.

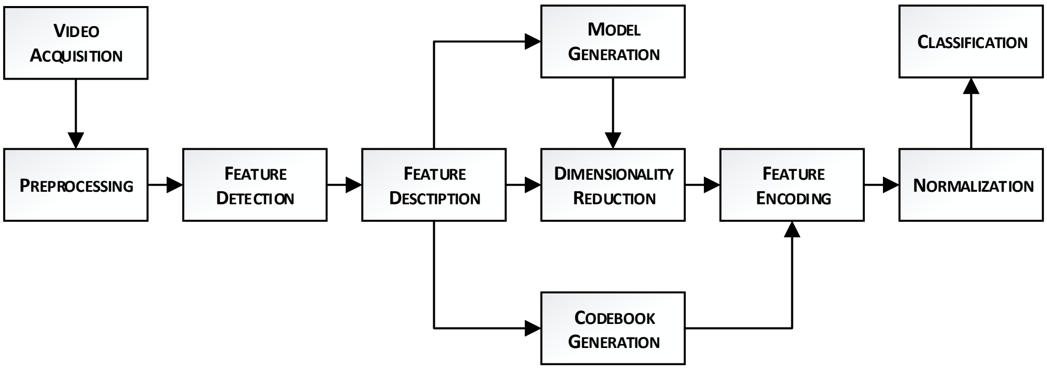

Figure 1. An illustration of the bag of visual words approach pipeline.

Table 7. Workflow of common elements and operations for human action recognition using spatio-temporal-based methods.

Workflow Elements	Variety of Operations
1. Feature Extraction:	(a) Trajectory-based features, (b) Dense Trajectories.
2. Motion Representation:	(a) Bag of Words, (b) RNNs, (c) CNNs, (d) Transformers.
3. Classification:	(a) SVM, (b) Random forest, (c) Neural Networks.

Pre-processing is optional. The videos can always be unified, e.g., rescaled or compressed. The feature detector chooses areas in the video that are volumes for computing features. Among the most popular are DI [186] and STIP [187], which provide sparse representation. However, using feature detectors is optional. Random sampling [188] and dense sampling [189] do not detect regions for feature extraction, which speeds up the final method. According to the current research, the dense sampling approach outperforms STIP [189]. Next, features within these sub-volumes are computed by the feature descriptor, popular descriptors include HOF, HOG [190], MBH [191], and fast GBH [188]. In the next phase, the dimensionality of the features is reduced by the popular PCA algorithm, which is a crucial element for performance [192]. Feature encoding clusters similar descriptors. Here, simple k-means and BoVW histogram or GMM and FV are utilized. These methods need a model and codebook to be established. The PCA model and codebook are usually learned from a subset of descriptors, e.g., as in [193]. The final representations are normalized and classified, usually by SVM with the RBF-χ^2 kernel for BoVW histogram descriptor and linear for FV. Many methods utilize different algorithms and combinations for each described phase of this approach. Some selected methods are presented below. This description is informative to analyse the method flowcharts in Appendix A. Unless stated otherwise, the methods mentioned use GMM with $k = 256$, FV, and SVM. The authors of the above cited examples often provide a study of different parameters. We take into account only the best reported results.

Heng Wang et al., in their work [189], compared the most popular descriptors, such as Cuboids, ESURF, HOF, HOG/HOF, HOG, and HOG3D, in combination with different detectors such as Harris3D, Cuboids, Hessian and Dense. The authors presented results on the following datasets: KTH, UCF, and Hollywood2. One of the essential conclusions for further research is that dense sampling detectors outperform sparse approaches. Based on the assumption of dense detectors, the dense trajectories method is a source approach for the best up-to-date methods that utilize hand-crafted features. For the most promising methods, flowcharts with the most critical blocks related to Figure 1 have been drawn and shown in Appendix A for visual comparison of the structures. First, the flowcharts for HAR by dense trajectories [194] and HAR with improved trajectories [195] are presented in

Figures A1 and A2. The one-against-rest approach is used in these multi-class classification cases, and the classes with the highest score are selected. The gradient boundary histograms for action recognition [188] pipeline in Figure A3 takes advantage of the random sampling method encoding local and the root channel separately. Pengs et.al, in their comprehensive study of BoVW methods [192], proposed the pipeline (Figure A4) composed of different BoVW methods and many different low-level descriptors. An efficient video representation and a robust approach for action recognition [196] combined iDT with spatial pyramid and spatial FV to preserve spatio-temporal features in the video presented in Figure A5. Beyond Gaussian pyramid: Multi-skip feature stacking for action recognition [197] (Figure A2) proposes efficient feature extraction at different time scales, encoding, and classification for action recognition similar to [198] (Figure A6). An efficient and effective human action recognition in the video through motion boundary description with a compact set of trajectories was presented in [199]. The method goes further with improved dense trajectories leading to better accuracy. The motion vector is interpolated between skipped frames to avoid computing optical flow and speed up the method. The following modification is that the number of trajectories per frame decreases below a threshold. A trajectory is also discarded in the case of too little motion within it. Fisher linear discriminant analysis (FLDA) is utilized for further dimensional reduction, working with sparse representation-based classification. Evaluation results demonstrate that there are fewer trajectories per frame than in iDT, and the methods are fast. Accuracy was also improved, but which change was crucial for this result was not evaluated. Action recognition with stacked fisher vectors [200] shows that SFV is effective in combination with standard FV. Here, the iDT is taken as the input, and a two-stage clustering structure is provided. This is a kind of mid-level approach without learning discriminative action parts. At the very beginning, a 396-dimensional descriptor is computed. It combines HOF, HOG, and MBH descriptors in sampled sub-cuboids. There are 600 to 6000 subcuboids, which differ across the datasets. Next, PCA and whitening reduce the dimension to 200. For each sub-cuboid, a consecutive FV is computed. This representation is, in turn, reduced by max-margin and further again by PCA and whitening, having 200 elements in the end. Another FV encodes the descriptors of a cuboid. Single-stage FV and SFVs are complementary. Its combination is one of the best available methods. Uijlings et al. [201] described popular descriptors in more detail and explained how to efficiently implement these algorithms to find a balance between accuracy and speed. The paper's authors [202] enriched video representation by focusing on encoding objects for actions and obtaining the best result by fusing FVs and SFVs and object-based proposed representation.

3.4. Deep Learning Methods

In modern DL architectures used for HAR, information concerning objects is usually included in video frames for the spatial and temporal dimensions of their movement. As shown in Table 8, the workflow for human action recognition using deep learning-based methods includes common elements and operations.

Table 8. Workflow of common elements and operations for human action recognition using deep learning-based methods.

Workflow Elements	Variety of Operations
1. Data Preparation:	(a) Data augmentation, (b) Pre-processing, (c) Data balancing.
2. Network Design:	(a) CNNs, (b) RNNs, (c) Transformers, (d) Attention mechanisms, (e) Hybrid networks.
3. Training:	(a) Backpropagation, (b) Regularization, (c) Optimization.

Since 2014, the most popular supervised DL model is the CNN, effectively applied for video HAR when Karpathy et al. [203] proposed a single-stream CNN to fuse temporal information from consecutive frames using pre-trained 2D convolutions. Later, Simonyan and Zisserman [204] presented a two-stream network architecture more suitable for the HAR task. The Simonyan method distinguishes temporal and spatial information using two separate streams for a CNN with three fully connected and five convolutional layers. The spatial part is trained on still images from the ImageNet challenge dataset [205]. The temporal part needs the stacking multiple-frame optical flow to be computed beforehand. The multi-task learning was performed on the most popular benchmarks UCF101 [178] and HMDB51 [177] datasets for the temporal part, and the accuracy of the computed soft-max scores were fused by linear SVM. These two papers form the backbone of most DL methods for HAR, differing in how spatio-temporal information is combined. Many other papers on single-, two- [206] or three-stream [207,208] architectures have evolved from these propositions. The most popular DL-based architectures applicable in HAR are presented in [209]. Due to their high computational complexity, multi-stream architectures are unsuitable for real-time surveillance applications where operators can adjust the system to new recognition classes. DL methods usually need a lot of computational time and data. They are challenging to analyse in detail, but selected variants with working code can compete with other methods in terms of accuracy. Some of the methods extend input from 2D performing 3D convolutions by 3D CNN with spatio-temporal information [210,211]. Combined methods use deep learning and hand-crafted features. Wang et al. in [212] combined the iDT approach with DL features. The best results were obtained by fusing DL descriptor with the traditional iDT approach at the FV level. A similar combination with improved FV (iFV) is presented in [213]. The VLAD [214] was used to encode spatio-temporal descriptors in combination with CNN [215–219]. Optical flow is a useful but inefficient motion model for CNN-based propositions, including two-stream [220,221] or faster modifications [222] and dynamic versions [223]. Some methods use CNN with skeleton sequences [224] to encode spatio-temporal information into texture patterns, others [225] use RGB-D representation for action scene flow. Temporal long-term relations are learned using sequential RNN [226] and LSTM [227–229] architectures. The workflow of common elements for human action recognition using deep learning-based methods, including data preparation, network design, and training, is presented in Table 8.

The idea of attention mechanism applied to computer vision [230] tries to estimate dependencies between relevant elements in consecutive video frames according to certain domains trying to learn the most important features or regions in an image or video by assigning different levels of attention to different parts of the input. According to [230], the channel attention mechanism (C) determines the importance of different channels (what to pay attention to), such as colour channels, in an RGB image. Spatial attention (H and W) determines the essential regions within an image based on their spatial location (where to pay attention). In contrast, temporal attention (T) is used to determine the critical frames in a video (when to pay attention). Branch attention combines these different attention mechanisms and provides a more comprehensive attention model.

These attention mechanisms are effective in various computer vision tasks, such as object detection, semantic segmentation, and video classification. By focusing on the most important features, attention mechanisms can help to improve the performance and efficiency of deep learning models in these tasks. Long et al. [231] applied attention to better capture temporal patterns in videos, and Dai et al. [232] proposed a spatio-temporal attention mechanism for feature learning processing to enhance the HAR performance.

The latest DL trend visual transformers (ViT) [233] could be a gamechanger in trying to parallelize operations by replacing the known drawbacks of sequential RNN architectures and, at the same time, limit the bias of locality from CNNs by using self-attention and two-stage training mechanisms. The main elements of ViT are presented in [234]. ViT's self-attention layer allows incorporating of global information throughout the entire image. To recreate the visual structure from the training data, ViT learns to encode the relative placement of the patches. Transformers lack prior knowledge of visual structure, resulting in increased training periods and the need for enormous datasets for model training. ViT separates the picture into visual tokens, whereas CNN employs pixel arrays. The video transformer network [235,236] for temporal relationships uses a long-former [237] to process the whole video in one pass. Action transformer networks try to aggregate spatio-temporal context cues around a selected person only using RGB frames [235]. Other propositions optimize the method of capturing spatio-temporal relations in videos [238,239]. Plizzari et al. [239] proposed a spatial self-attention module and temporal self-attention transformer for inter-frame correlations to model the human skeleton structure.

Despite the exceptional performance of transformer models for standard HAR benchmarks and intriguing prominent features, there are significant problems related to their practical use. The demand for enormous volumes of training data and the highest computing costs are the most significant barriers. Visualizing and interpreting transformer models has also proven challenging. We present a summary of these problems in this part, along with some recent initiatives to overcome these constraints.

Transformers provide an easy way to see what they are paying attention to [240], while this does not give a complete indication of the types of associations learned by the model [241], it does provide some insight into what it considers significant for specific samples [242]. Few studies have attempted to interpret transformers further than this for vision [243]. We only identified a small portion of research depicting these attention activations for individual samples in the ViT literature.

3.5. Datasets for Method Evaluation

There are always complex problems to solve in videos from a surveillance camera, such as changing light conditions, background clutter, and occlusions. Numerous datasets are available for benchmarking and comparing human action recognition methods [244]. The most up-to-date paper [245], published in May of 2022, presents an excellent vast summary with a catalog of the 704 existing multimodal human movement datasets available for researchers prepared in labs and the real world. Table 9 indicates the most popular datasets from real-life scenarios considered when selecting the most representative datasets to evaluate the most promising methods. The state-of-the-art methods often use HMDB51, Hollywood2, and UCF101 for benchmarking. We have followed this direction and extended this set of benchmarks with one additional test with the VMASS2 dataset, where all video streams come from surveillance camera networks in the metropolitan area. Several publicly available datasets have also been widely used in the research community to evaluate the performance of IVA algorithms. Some of the most popular datasets: RWF-2000, XD-Violence, and UCF-Crime, include videos with registered violence. Weizmann and KTH are datasets of human actions and objects captured using a static camera. Weizmann was performed by nine people and KTH by 25 actors. The VMASS dataset includes a diverse range of human actions captured from various surveillance camera angles under different lighting conditions, making it a challenging and comprehensive dataset to evaluate the performance of IVA algorithms. On the other hand, the UCF sport action and UCF11 datasets consist of human actions captured from YouTube videos. The UCF101 dataset includes 101 different human activities, while the HMDB51 dataset contains 51. The key features of the VMASS dataset include its large scale, diverse action categories, and multimodal annotations, which provide a rich resource for developing and evaluating new IVA algorithms. Each of these datasets include a diverse range of human actions recorded using various cameras and under different conditions, making them well-suited for assessing the robustness and accuracy of IVA algorithms.

Table 9. Selected datasets for benchmarking human action recognition methods from real-life scenarios.

Name	Videos	Classes
RWF-2000 [183]	2000	2
KTH [246]	2391	6
XD-Violence [247]	4754	9
Weizmann [248]	90	10
UCF sport action [249]	150	10
UCF11 (YouTube) [250]	1160	11
Hollywood2 [179]	1707	12
UCF-Crime [251]	1900	13
Olympic Sports [252]	783	16
UCF50 [253]	6676	50
HMDB51 [177]	6849	51
MultiTHUMOS [254]	400	65
UCF101 [178]	>13,000	101
NTU RGB+D 120 [255]	114,000	120
VMASS2 [180]	>6,000,000	400
Kinetics 700 [256]	650,000	700

4. Discussion

In this paper, we briefly presented the state of knowledge of modern IVA architectures, which has become a generally available trend in video surveillance systems in recent years.

4.1. IVA Systems

Despite the shortcomings of advanced IVA technologies, most systems struggle with the problems of business continuity, efficient alerting and response, and the inability to dynamically track a detected event in the camera network. Despite technological advancements, the current systems do not have modules to effectively recognize the actions and behaviours of a broad spectrum of events and scenes observed under various lighting and weather conditions. In addition, the systems available on the market do not have modules that allow operators to train systems to learn events directly from the video stream. Each function related to adapting the existing video surveillance system involves a tedious process of collecting specific data, developing new models based on it, and then implementing them into the existing infrastructure. Due to the requirements of many algorithms, such implementation often forces companies to replace the existing computing equipment with new ones to support computationally demanding algorithms. For the needs of distributed surveillance systems, the VSaaS service has been introduced, which allows the customer's attention to focus on specific areas or events that interest them. The provider of such a service bears the equipment costs and maintenance in this variant created on-demand in a distributed environment of many connected cameras or other multimodal devices forming local surveillance systems while simultaneously being part of the global network of the IoT. These local infrastructures are parts of a more extensive ecosystem, causing an even greater demand for algorithms and services to increase their situational awareness of the monitored sites.

The main research directions regarding recognizing people's actions from a video stream have been presented. These include pose estimation, tracking, deep learning, and space-time-based methods. This part summarizes each of the directions listed above.

4.2. Tracking-Based Methods

The current generation of visual trackers has a problem with scene understanding. Existing approaches cannot detect global structures and existent objects or interpret dynamic circumstances meaningfully. In this few-data regime scenario, newer trackers based on adversarial learning may be an alternative and include additional attributes such as spatiotemporal information.

The fundamental goal of modern tracking methods is to create unique neural networks that are simultaneously resilient, accurate, and efficient. Most recent studies have not pretrained or fine-tuned their backbone networks to utilize generic characteristics and prevent catastrophic forgetting of general patterns. Researchers suggest various machine learning-based strategies to overcome this issue and have demonstrated by preliminary works that adequate backbone network training can improve tracking performance.

Despite significant developments in short-term trackers, long-term trackers are disregarded. On the other hand, long-term trackers seem more useful in real-world circumstances where the tracking object may often disappear or remain occluded for an extended time. After a failure, these trackers should be able to detect the tracking object again and then continue monitoring the proper object in the video stream.

The modern direction—deep learning-based visual tracking approaches have recently examined various uses of deep features, a fusion of hand-crafted and deep features, search strategies, various topologies, and training on datasets, and how to cope with missing training data. However, these are not stable solutions, and many difficulties remain unsolved, and others will need to be investigated further in the future.

Existing tracking-based methods may struggle with tracking multiple people simultaneously, especially when they are close together or appear similarly. Multi-person tracking is a challenging problem in HAR that requires developing effective methods to handle occlusion, appearance changes, and interactions between individuals.

4.3. Pose-Based Methods

Despite the promising results, some 2D human pose recognition problems still need to be solved in future studies. Processing efficiency is one of the known problems. Specific frameworks, such as OpenPose, may accomplish near real-time processing on dedicated hardware with a moderate computational capacity. However, more efficient human pose estimation techniques on commercial devices are required in real-world applications.

Another issue is the shortage of data for unusual positions. Whereas existing 2D human pose estimation datasets are big enough for traditional postures, they have inadequate training data for unexpected poses, such as fighting. Model bias and poor performance in unique postures may occur from data imbalance.

The next problem concerns recognizing a person in crowded and natural situations with multiple bodies and other objects occluded. Person detectors may miss the borders of highly overlapping human bodies. In occluded situations, the difficulty of keypoint association is also more evident for bottom-up techniques.

Model generalization is one of the challenges for 3D pose-based methods. Motion capture systems are a bottleneck because they require high-quality 3D ground truth posture annotations, which are expensive and difficult to install in a random environment. As a result, most current datasets have been collected from confined scenarios. On these datasets, state-of-the-art algorithms produce promising results, but their performance declines when applied to real-world data.

The 3D human pose estimation requires substantially more computation than 2D estimation. It is challenging to develop computationally efficient 2D human posture estimate pipelines while keeping high pose estimation accuracy. Due to extreme mutual occlusions and poor resolution content of each individual, the performance of existing 3D human pose estimation algorithms suffer significantly in crowded scenes. Nevertheless, the critical findings are worth discussing since pose-based techniques are one of HAR's most promising avenues.

4.4. Deep Learning-Based Methods

Neural networks are continuously becoming more advanced and are often applied to computer vision problems such as HAR. Some of the modern methods from the above mentioned research also use elements of deep learning—the most studied and utilized methods are CNN and, more recently, ViT. Since ViT is much more advantageous than CNN, we have listed the common disadvantages of these methods that should be considered when selecting these methods.

Transformer models are known for their ability to scale to high levels of parametric complexity, while this is a fantastic trait that enables the formation of massive models, it comes at a hefty cost in training and inference, e.g., according to estimates. The process of training the GPT3 model with 175 billion parameters might cost OpenAI USD 4.6 million. The high computing cost of transformer models also affects computer vision models. Image generators based on sequence-based transformers (such as iGPT) have a high computation cost, limiting their application to high-resolution inputs. In transformers, the time and memory cost of the fundamental self-attention process grows quadratically with the number of image patches.

Transformer designs often require a lot of training to determine the underlying modality-specific principles because they do not natively incorporate prior knowledge to deal with the visual input. The self-attention system must automatically uncover relationships between video frames by analysing an extensive library of video sequences. This process leads to lengthier training durations, higher computational needs, and the processing of big datasets. To achieve a decent performance on the ImageNet benchmark dataset, the ViT [257] model, for example, requires hundreds of millions of pictures. The difficulty of training a transformer in a data-efficient manner is still an open research subject, although recent studies show promising progress.

These significant drawbacks make this direction promising but not mature enough for practical application and research. Nevertheless, the topic of DL to recognize actions in a video is pervasive. There are many comprehensive reviews of the use of classic DL methods to identify human actions [258] as well as video transformers applied to computer vision tasks [233,234,259].

5. Conclusions

This paper comprehensively reviews existing human action recognition methods for intelligent video analytics. We examined the advantages and disadvantages of spatio-temporal, pose-based, tracking-based, and deep learning-based approaches, as well as the potential applications of each. Spatio-temporal methods use motion information to capture action patterns, while pose-based techniques utilize body posture to identify human actions. Tracking-based methods use tracking algorithms to identify action sequences, and deep learning-based methods utilize neural networks to classify human activities. Additionally, we compared classic and edge AI intelligent video analytics systems in the cloud, on-premises, and on edge. Popular edge AI neural systems such as Google Coral, Intel Myriad, and Kneros are increasingly used for human action recognition. Google Coral is a system-on-module based on a low-power edge TPU chip, Intel Myriad X is designed for computer vision at the edge, and Kneros is an AI-enabled system-on-module. Deep learning models can be deployed on such neural systems using a wide variety of pre-trained models from the Model Zoo, or custom models can be built and trained with the help of AutoML. Classic AI systems are typically hosted in the cloud, while edge AI systems are designed to run locally on-premises or at the network's edge. Furthermore, this paper outlined the challenges and opportunities of human action recognition in intelligent video analytics, suggesting possible future research directions. We also provided an in-depth analysis of important aspects of current methods and their potential to improve smart video analytics. Due to the previously mentioned shortcomings of the stability and performance of the presented methods, it is tough to choose one particular class of HAR methods and develop a comprehensive surveillance system enabling the training and real-time recognition of moving objects in broad-spectrum weather conditions. The modern deep learning-based HAR methods show the most promising results but in limited cases. The newest variants of these methods often have high algorithmic and overall computational complexity.

To construct a commercial system to recognize actions from a video stream, each activity in the video data processing pipeline should be explainable, avoiding the unexpected operation of algorithms, and stable, predictable architecture should simplify utilization and allow regular system maintenance and upgrades without the necessity of replacing the current system with an entirely new one.

This review serves as a guide for researchers and practitioners to better understand the last 20 years of research, the current state-of-the-art human action recognition technologies for intelligent video analytics and identify potential opportunities for future research. One classic group of spatio-temporal methods is based on stable, known, and simple spatial and temporal feature detection and description algorithms. Most methods from the spatio-temporal group proved to work fast and predictably with real-world video data in many practical applications. The class of BoVW methods seems to be more robust to environmental changes since they rely on the appearance of objects rather than their spatial relationships. Additionally, BoVW techniques provide better generalization capabilities, as they are less likely to overfit when presented with new data. Finally, BoVW methods are easier to implement and require less training data than other models. Therefore, BoVW methods are good candidates as human action recognition modules in intelligent video analytics before ViT-based methods become more mature with a cheaper entry point. Visual transformers (ViTs) are a more recent approach to image recognition, and have already shown promising results in various computer vision tasks. Unlike BoVW, ViTs rely on self-attention mechanisms to capture the relationships between image features without the need for explicit spatial binning or pooling. This allows ViTs to model more complex and abstract relationships between features, making them more suitable for tasks that require a higher level of understanding of the input images. However, ViTs currently require a large amount of training data and computational resources to achieve a state-of-the-art performance, a major limitation in some applications. In contrast, BoVW methods are relatively simple to implement and require less data to train, making them more suitable for applications with limited data and computational resources.

Overall, both BoVW and ViT approaches have their strengths and weaknesses, and the choice between them depends on the specific requirements and constraints of the application. It is worth remembering that each method has unique characteristics and strengths. A comprehensive approach to analysing human movement may involve combining these methods depending on the specific task.

Author Contributions: Conceptualization, M.K. (Marek Kulbacki) and J.S.; validation, J.W.R., Z.C. and K.W.; investigation, M.K. (Marek Kulbacki); data curation, M.K. (Marek Kulbacki); writing—original draft preparation, M.K. (Marek Kulbacki); writing—review and editing, M.K. (Marek Kulbacki), J.S., Z.C., J.W.R., M.K. (Michał Kulbacki), R.K. and K.W.; visualization, M.K. (Michał Kulbacki); supervision, M.K. (Marek Kulbacki). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

IVA	Intelligent video analytics
CCTV	Closed circuit television
CCD	Charge-coupled device
HAR	Human activity recognition
VSaaS	Video surveillance as a service
BoVW	Bag of visual words
RoI	Region of interest

Appendix A. Bag of Visual Words—Promising Classic Representations

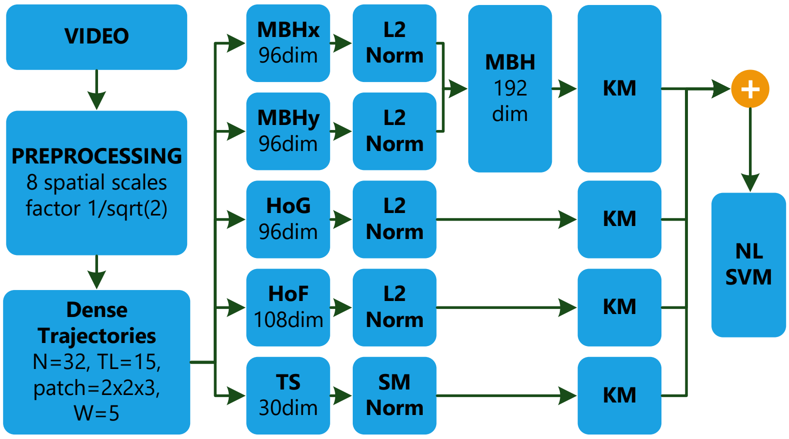

Figure A1. An illustration of the dense trajectories method [194]. TS : Trajectory shapes; SM Norm: Norm by Sum of Magnitudes; KM: K-means with $K = 4000$, 100,000 randomly selected features performed eight times, selected result with the lowest error; NL SVM: Non-linear SVM with the RBF-χ^2 kernel.

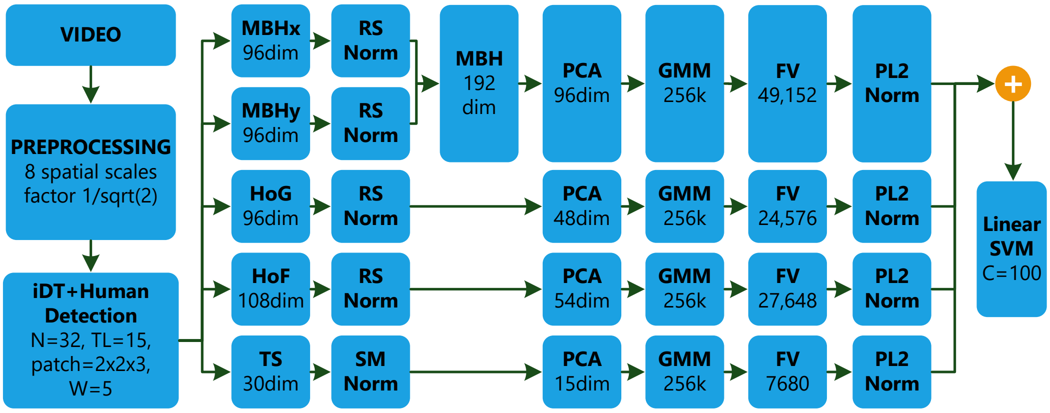

Figure A2. This figure illustrates the backbone for two different approaches to HAR. Variant 1 uses the improved dense trajectories method [195], which includes Trajectory Shapes (TS), Norm by Sum of Magnitudes (SM Norm), Gaussian Mixture Model (GMM), Root SIFT Norm (RS Norm), Fisher Vector (FV), and Power and L2 Norm (PL2 Norm). Variant 2 is an extended method that builds on Variant 1 and incorporates multi-skip feature stacking [197]. It uses iDT + Human Detection to extract descriptors and skips frames to reduce the number of frames processed. Skipping each 2nd frame adds 0.5 more frames and descriptors, equivalent to having a 0.5 longer video, while skipping each 3rd frame adds 0.3 more frames, and skipping each 4th frame adds 0.25 more frames. Variant 2 includes RS Norm, TS, SM Norm, GMM, FV, and PL2 Norm.

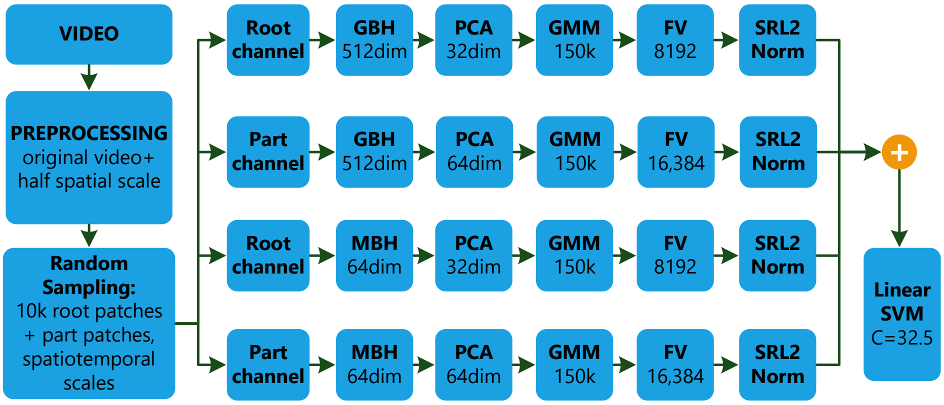

Figure A3. An illustration of the gradient boundary histogram method [188]. Random Sampling: A total of 10,000 root patches at half spatial resolution, eight (2 × 2 × 2) overlapping part patches. Eight spatial and two temporal scales. Initial patch size 28 × 28 × 14. Each patch is subdivided into 2 × 2 × 2 cells, which together with eight bins gives 64 dim descriptor TS: Trajectory shapes; SM Norm: Norm by Sum of magnitudes; GMM: Gaussian Mixture Model; FV: Fisher Vector; SRL2 Norm: Square Rooting and L2 norm.

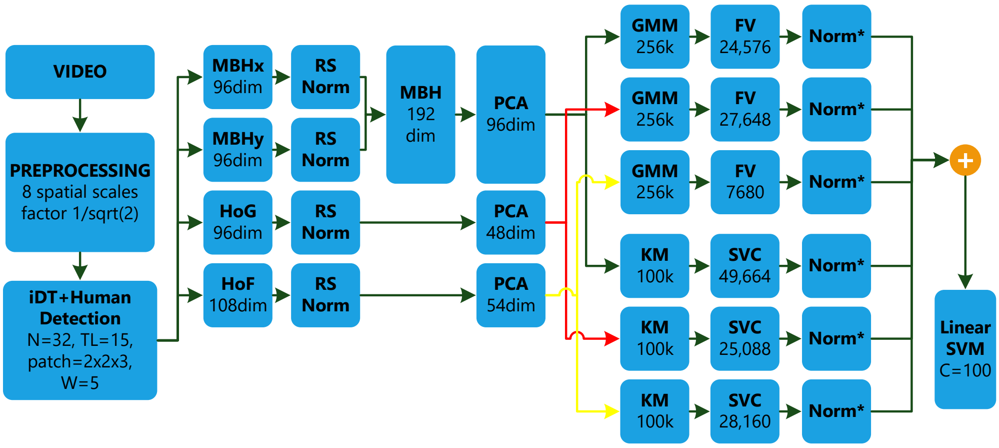

Figure A4. An illustration of the hybrid Peng's system [192]. RS Norm: Root SIFT norm; SM Norm: Norm by Sum of Magnitudes; GMM: Gaussian Mixture Model; FV: Fisher Vector; Norm*: Power Norm with factor α = 0.5 which is signed square rooting and then intra L2 Norm.

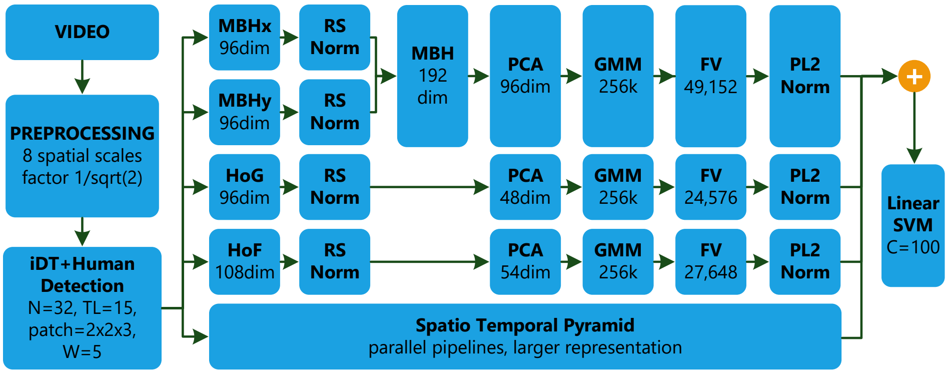

Figure A5. An illustration of the robust and efficient INRIA method [196]. Additional annotations with video, bounding boxes with humans, bring improvement. RS Norm: Root SIFT Norm; Spatio-Temporal Pyramid: Parallel the same steps in smaller cuboids. H3 enlarges the dimension four times, T2 enlarges it three times T2 + H3 enlarges the dimension six times; GMM: Gaussian Mixture Model; FV: Fisher Vector; PL2 Norm: Power Norm with factor α = 0.5 which is signed Square Rooting and then L2 Norm.

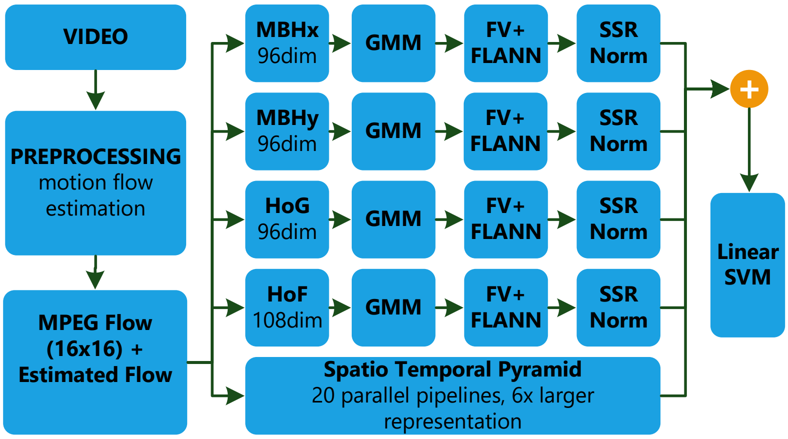

Figure A6. An illustration of the efficient feature extraction, encoding, and classification method for AR [198]. MPEG compression and decompression for estimation of motion flow (used instead of optical flow) during pre-processing step. MPEG flow (16 × 16) + estimated flow (to increase the spatial resolution by a factor of 2). A very comprehensive spatio-temporal pyramid; a total of 24 cells, two scales: 32 × 32 × 15 with spatial stride 16 and temporal stride 548 × 48 × 15 with spatial stride 24 and temporal stride 5 descriptor's area: 2 × 2 × 3; GMM: Gaussian Mixture Model; FV: Fisher Vector; FLANN: Fast Library for Approximate Nearest Neighbors SSR Norm: Signed Square Root (power normalization with an α factor of 0.5) and L2 Norm.

References

1. Research, D. *Global Surveillance Camera Market: Analysis by System Type (Analog, IP Commercial, IP Consumer and Other Surveillance Camera), by Technology (Image Signal Processor, Vision Processor, Vision Processor + AI) by Region Size and Trends with Impact of COVID-19 and Forecast up to 2027*; Technical Report DAR17374302; Daedal Research: Dublin, Ireland, 2022.
2. Davis, L.S. *Real Time Computer Surveillance for Crime Detection*; Technical Report; University of Maryland: College Park, MD, USA, 2001.
3. Lyon, D. *Surveillance Studies: An Overview*; Polity: Cambridge, UK, 2007.
4. Ratcliffe, J. *Response Guides Series Problem-Oriented Guides for Police Video Surveillance of Public Places*; Center for Problem-Oriented Policing, Inc.: Phoenix, AZ, USA, 2011.
5. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S. A review of video surveillance systems. *J. Visual Commun. Image Represent.* **2021**, *77*, 103116. [CrossRef]
6. Hamoudy, M.A.; Qutqut, M.H.; Almasalha, F. Video security in Internet of things: An overview. *Int. J. Comput. Sci. Netw. Secur.* **2017**, *17*, 199.
7. Volker, E.; Töpfer, E. The Human- and Hardware of Policing Neoliberal Sport Events: Rent-a-Cops, Volunteers and CCTV at the FIFA Championship in Germany 2006—And beyond. In Proceedings of the Conference Security and Surveillance at Mega Sport Events, Durham University, Durham, UK, 4–8 April 2008; Volume 25.
8. King, J.; Mulligan, D.K.; Raphael, S.P. CITRIS Report: The San Francisco Community Safety Camera Program—An Evaluation of the Effectiveness of San Francisco's Community Safety Cameras. *SSRN Electron. J.* **2008**. [CrossRef]
9. Deisman, W.; Derby, P.; Doyle, A.; Leman-Langlois, S.; Lippert, R.; Lyon, D.; Pridmore, J.; Smith, E.; Walby, K.; Whitson, J. A Report on Camera Surveillance in Canada Part One Surveillance Camera Awareness Network (SCAN). In *Surveillance Project: Surveillance Camera Awareness Network (SCAN)*; Social Sciences and Humanities Research Council: Ottawa, ON, Canada, 2009.
10. Runolfson, D.; Intern, A. *Cal Anderson Park Surveillance Camera Pilot Program Evaluation*; Technical Report; The Office of City Auditor: Seattle, WA, USA, 2009.
11. Hempel, L.; Töpfer, E. *CCTV in Europe*; Final Report; Centre for Technology and Society Technical University Berlin: Berlin, Germany, 2004; Volume 15. Available online: http://www.urbaneye.net/results/ue_wp15.pdf (accessed on 2 March 2023).
12. Newell, B.C. (Ed.) *Police on Camera: Surveillance, Privacy, and Accountability*, 1st ed.; Routledge Studies in Surveillance; Routledge: Avenue, NY, USA, 2020.
13. Park, Y.J. *The Future of Digital Surveillance: Why Digital Monitoring Will Never Lose Its Appeal in a World of Algorithm-Driven AI*; University of Michigan Press: Ann Arbor, MI, USA, 2021.
14. Brown, L.; Hampapur, A.; Connell, J.; Lu, M.; Senior, A.; Shu, C.F.; Tian, Y. *IBM Smart Surveillance System (S3): An Open and Extensible Architecture for Smart Video Surveillance*; In Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, Como, Italy, 15–16 September 2005.
15. BenAbdelkader, C.; Burlina, P.; Davis, L. *Gait as a Biometric for Person Identification in Video Sequences*; Technical Report; University of Maryland: College Park, MD, USA, 2001.

16. Dick, A.R.; Brooks, M.J. Issues in Automated Visual Surveillance. In Proceedings of the International Conference on Digital Image Computing: Techniques and Applications, Sydney, Australia, 10–12 December 2003; Sun, C., Talbot, H., Ourselin, S., Adriaansen, T., Eds.; CSIRO Publishing: Clayton, Australia, 2003; pp. 195–204.
17. Filipi Gonçalves dos Santos, C.; Oliveira, D.d.S.; Passos, L.A.; Gonçalves Pires, R.; Felipe Silva Santos, D.; Pascotti Valem, L.; Moreira, T.P.; Santana, M.C.S.; Roder, M.; Paulo Papa, J.; et al. Gait Recognition Based on Deep Learning: A Survey. *ACM Comput. Surv.* **2022**, *55*, 3490235. [CrossRef]
18. Ko, T. A survey on behavior analysis in video surveillance for homeland security applications. In Proceedings of the 2008 37th IEEE Applied Imagery Pattern Recognition Workshop, Washington, DC, USA, 15–17 October 2008. [CrossRef]
19. Collins, R.T.; Lipton, A.J.; Kanade, T.; Fujiyoshi, H.; Duggins, D.; Tsin, Y.; Tolliver, D.; Enomoto, N.; Hasegawa, O.; Burt, P.; et al. *A System for Video Surveillance and Monitoring—CMU-RI-TR-00-12*; Technical Report; Carnegie Mellon University: Pittsburgh, PA, USA, 2000.
20. Shankar, G.; Latchoumi, T.; Chithambarathanu, M.; Balayesu, N.; Shanmugapriya, C. An Efficient Survey on Energy Conservation System with Video Surveillance. *J. Xian Univ. Archit. Technol.* **2020**, *12*, 106.
21. Borg, M.; Thirde, D.; Ferryman, J.; Florent, F.; Valentin, V.; Brémond, F.; Thonnat, M. Video Surveillance for Aircraft Activity Monitoring. In Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, Como, Italy, 15–16 September 2005; pp. 16–25. [CrossRef]
22. Ferryman, J.; Shahrokni, A. Pets2009: Dataset and challenge. In Proceedings of the 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Snowbird, UT, USA, 7–12 December 2009; pp. 1–6. [CrossRef]
23. Brémond, F.; Thonnat, M.; Zúniga, M.Z. Video-understanding framework for automatic behavior recognition. *Behav. Res. Methods* **2006**, *38*, 416–426. [CrossRef]
24. Vincent, P.; Driver, M.; Wang, J. *Low-Code Development Technologies Evaluation Guide*; Technical Report; Gartner Research: Stamford, CT, USA, 2019.
25. Wang, L.; Hu, W.; Tan, T. Recent developments in human motion analysis. *Pattern Recognit.* **2003**, *36*, 585–601. [CrossRef]
26. Hu, W.; Tan, T.; Wang, L.; Maybank, S. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2004**, *34*, 334–352. [CrossRef]
27. 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. [CrossRef]
28. Iguernaissi, R.; Merad, D.; Aziz, K.; Drap, P. People tracking in multi-camera systems: A review. *Multimedia Tools Appl.* **2019**, *78*, 10773–10793. [CrossRef]
29. Poppe, R. Vision-based human motion analysis: An overview. *Comput. Vis. Image Underst.* **2007**, *108*, 4–18. [CrossRef]
30. Kumar, P.; Mittal, A.; Kumar, P. Study of Robust and Intelligent Surveillance in Visible and Multi-modal Framework. *Informatica* **2008**, *32*, 63–77.
31. Antonakaki, P.; Kosmopoulos, D.; Perantonis, S.J. Detecting abnormal human behaviour using multiple cameras. *Signal Process.* **2009**, *89*, 1723–1738. [CrossRef]
32. Brand, M.; Kettnaker, V. Discovery and segmentation of activities in video. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 844–851. [CrossRef]
33. Stauffer, C.; Grimson, W.E. Adaptive background mixture models for real-time tracking. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Fort Collins, CO, USA, 23–25 June 1999; Volume 2, pp. 246–252. [CrossRef]
34. Alipour, P.; Shahbahrami, A. An adaptive background subtraction approach based on frame differences in video surveillance. In Proceedings of the 2022 International Conference on Machine Vision and Image Processing (MVIP), Ahvaz, Iran, 23–24 February 2022; pp. 1–5. [CrossRef]
35. Shah, S.T.H.; Xuezhi, X. Traditional and modern strategies for optical flow: An investigation. *SN Appl. Sci.* **2021**, *3*, 1–14. [CrossRef]
36. Alzughaibi, A.; Chaczko, Z. Human Detection Using Illumination Invariant Feature Extraction for Natural Scenes in Big Data Video Frames. In Proceedings of the 2017 25th International Conference on Systems Engineering (ICSEng), Las Vegas, NV, USA, 22–23 August 2017; pp. 443–450. [CrossRef]
37. Huang, Z.; Shi, X.; Zhang, C.; Wang, Q.; Cheung, K.C.; Qin, H.; Dai, J.; Li, H. FlowFormer: A Transformer Architecture for Optical Flow. *arXiv* **2022**, arXiv:2203.16194.
38. Shi, H.; Zhou, Y.; Yang, K.; Ye, Y.; Yin, X.; Yin, Z.; Meng, S.; Wang, K. PanoFlow: Learning optical flow for panoramic images. *arXiv* **2022**, arXiv:2202.13388.
39. Bobick, A.F.; Davis, J.W. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 257–267. [CrossRef]
40. Segen, J.; Kumar, S. Look Ma, No Mouse. Human-Computer Interaction Using Hand Gestures. *Commun. ACM* **2000**, *43*, 102–109. [CrossRef]
41. Ivanov, Y.A.; Bobick, A.F. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 852–872. [CrossRef]
42. Segen, J.; Pingali, S.G. A camera-based system for tracking people in real time. In Proceedings of the 13th International Conference on Pattern Recognition, Vienna, Austria, 25–29 August 1996; Volume 3, pp. 63–67. [CrossRef]

43. Segen, J.; Pingali, S. An Inductive System for Tracking People in Live Video. In Proceedings of the IEEE Workshop on Machines that Learn, Stockholm, Sweden, 18–20 June 1996.
44. Cavallaro, A.; Steiger, O.; Ebrahimi, T. Tracking video objects in cluttered background. *IEEE Trans. Circuits Syst. Video Technol.* **2005**, *15*, 575–584. [CrossRef]
45. Javed, O.; Shah, M. Tracking and Object Classification for Automated Surveillance. In *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part IV 7*; Springer: Berlin/Heidelberg, Germany, 2002; Volume 2353, pp. 343–357. [CrossRef]
46. Isard, M.; Blake, A. Contour tracking by stochastic propagation of conditional density. In *Computer Vision—ECCV'96: 4th European Conference on Computer Vision Cambridge, UK, April 15–18, 1996 Proceedings, Volume I 4*; Springer: Berlin/Heidelberg, Germany, 1996; Volume 1064, pp. 343–356. [CrossRef]
47. Alzughaibi, A.; Chaczko, Z. Human detection model using feature extraction method in video frames. In Proceedings of the 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ), Palmerston North, New Zealand, 21–22 November 2016; pp. 1–6. [CrossRef]
48. Doucet, A.; de Freitas, N.; Gordon, N. *Sequential Monte Carlo Methods in Practice*; Springer: New York, NY, USA, 2001. [CrossRef]
49. Isard, M.; Blake, A. Condensation—Conditional Density Propagation for Visual Tracking. *Int. J. Comput. Vis.* **1998**, *29*, 5–28. [CrossRef]
50. Bregler, C. Learning and Recognizing Human Dynamics in Video Sequences. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 17–19 June 1997; p. 568. [CrossRef]
51. Medioni, G.G.; Cohen, I.; Brémond, F.; Hongeng, S.; Nevatia, R. Event Detection and Analysis from Video Streams. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 873–889. [CrossRef]
52. Segen, J.; Pingali, S. Video Based Tracking and Analysis of Human Movements. In Proceedings of the CVPR'96, San Francisco, CA, USA, 18–20 June 1996.
53. Pingali, G.; Segen, J. People Tracking for Automatic Identification. In *IEEE Workshop on Advanced Automatic Identification Technologies*; IEEE: Piscataway, NJ, USA, 1997.
54. Cèdras, C.; Shah, M. Motion-based recognition a survey. *Image Vis. Comput.* **1995**, *13*, 129–155. [CrossRef]
55. Koller-Meier, E.B.; Gool, L.V. Modeling and Recognition of Human Actions Using a Stochastic Approach. In *Video-Based Surveillance Systems*; Springer: Boston, MA, USA, 2002; pp. 179–191. [CrossRef]
56. Makris, D.; Ellis, T. Learning semantic scene models from observing activity in visual surveillance. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2005**, *35*, 397–408. [CrossRef] [PubMed]
57. Bobick, A.F.; Wilson, A.D. A State-Based Approach to the Representation and Recognition of Gesture. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 1325–1337. [CrossRef]
58. Jan, T. Neural network based threat assessment for automated visual surveillance. In Proceedings of the IEEE International Conference on Neural Networks, Budapest, Hungary, 25–29 July 2004; Volume 2, pp. 1309–1312. [CrossRef]
59. Smith, M.I.; Heather, J.P. A review of image fusion technology in 2005. *Thermosense XXVII* **2005**, *5782*, 29–45. [CrossRef]
60. Heartwell, C.H.; Lipton, A.J. Critical asset protection, perimeter monitoring and threat detection using automated video surveillance—A technology overview with case studies. In Proceedings of the IEEE Annual International Carnahan Conference on Security Technologys, Atlantic City, NJ, USA, 24–24 October 2002; p. 87. [CrossRef]
61. Szarvas, M.; Sakait, U.; Ogata, J. Real-time pedestrian detection using LIDAR and convolutional neural networks. In Proceedings of the IEEE Intelligent Vehicles Symposium, Meguro-Ku, Japan, 13–15 June 2006; pp. 213–218. [CrossRef]
62. Premebida, C.; Monteiro, G.; Nunes, U.; Peixoto, P. A Lidar and vision-based approach for pedestrian and vehicle detection and tracking. In Proceedings of the IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, Bellevue, WA, USA, 30 September–3 October 2007; pp. 1044–1049. [CrossRef]
63. Morris, B.T.; Trivedi, M.M. A survey of vision-based trajectory learning and analysis for surveillance. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1114–1127. [CrossRef]
64. Heilbron, F.C.; Escorcia, V.; Ghanem, B.; Niebles, J.C. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 961–970. [CrossRef]
65. Kong, Q.; Wu, Z.; Deng, Z.; Klinkigt, M.; Tong, B.; Murakami, T. MMAct: A Large-Scale Dataset for Cross Modal Human Action Understanding. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019. [CrossRef]
66. Lavee, G.; Rivlin, E.; Rudzsky, M. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2009**, *39*, 489–504. [CrossRef]
67. Hamid, R.; Maddi, S.; Johnson, A.; Bobick, A.; Essa, I.; Isbell, C. A novel sequence representation for unsupervised analysis of human activities. *Artif. Intell.* **2009**, *173*, 1221–1244. [CrossRef]
68. Wang, X. Intelligent multi-camera video surveillance: A review. *Pattern Recognit. Lett.* **2013**, *34*, 3–19. [CrossRef]
69. Chen, J.; Ran, X. Deep learning with edge computing: A review. *Proc. IEEE* **2019**, *107*, 1655–1674. [CrossRef]
70. Chen, T.; Li, M.; Li, Y.; Lin, M.; Wang, N.; Wang, M.; Xiao, T.; Xu, B.; Zhang, C.; Zhang, Z. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv* **2015**, arXiv:1512.01274.

71. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. {TensorFlow}: A System for {Large-Scale} Machine Learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
72. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678. [CrossRef]
73. Foundation, T.L. *State of the Edge Report*; Linux Foundation: San Francisco, CA, USA, 2021.
74. Bilal, K.; Khalid, O.; Erbad, A.; Khan, S.U. Potentials, trends, and prospects in edge technologies: Fog, cloudlet, mobile edge, and micro data centers. *Comput. Netw.* **2018**, *130*, 94–120. [CrossRef]
75. Gavrila, D. The Visual Analysis of Human Movement: A Survey. *Comput. Vis. Image Underst.* **1999**, *73*, 82–98. [CrossRef]
76. Aggarwal, J.K.; Ryoo, M.S. Human activity analysis: A review. *ACM Comput. Surv. (CSUR)* **2011**, *43*, 16. [CrossRef]
77. Negin, F.; Bremond, F. *Human Action Recognition in Videos: A Survey*; INRIA Technical Report; INRIA: Paris, France, 2016.
78. Onofri, L.; Soda, P.; Pechenizkiy, M.; Iannello, G. A survey on using domain and contextual knowledge for human activity recognition in video streams. *Expert Syst. Appl.* **2016**, *63*, 97–111. [CrossRef]
79. Herath, S.; HAR-Surveyandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **2017**, *60*, 4–21. [CrossRef]
80. Wu, D.; Sharma, N.; Blumenstein, M. Recent advances in video-based human action recognition using deep learning: A review. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017. [CrossRef]
81. 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]
82. Zhang, H.B.; Zhang, Y.X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.X.; Chen, D.S. A comprehensive survey of vision-based human action recognition methods. *Sensors* **2019**, *19*, 1005. [CrossRef]
83. Kong, Y.; Fu, Y. Human action recognition and prediction: A survey. *Int. J. Comput. Vis.* **2022**, *130*, 1366–1401. [CrossRef]
84. Chakraborty, S.; Mondal, R.; Singh, P.K.; Sarkar, R.; Bhattacharjee, D. Transfer learning with fine tuning for human action recognition from still images. *Multimedia Tools Appl.* **2021**, *80*, 20547–20578. [CrossRef]
85. Naqushbandi, F.S.; John, A. Sequence of actions recognition using continual learning. In Proceedings of the 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India, 23–25 February 2022; pp. 858–863. [CrossRef]
86. Wang, C.; Qiu, Y.; Gao, D.; Scherer, S. Lifelong Graph Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–20 June 2022; pp. 13719–13728.
87. Xiao, Z.; Xu, X.; Xing, H.; Song, F.; Wang, X.; Zhao, B. A federated learning system with enhanced feature extraction for human activity recognition. *Knowl.-Based Syst.* **2021**, *229*, 107338. [CrossRef]
88. Hegedűs, I.; Danner, G.; Jelasity, M. Decentralized learning works: An empirical comparison of gossip learning and federated learning. *J. Parallel Distrib. Comput.* **2021**, *148*, 109–124. [CrossRef]
89. Zhu, R.; Xiao, Z.; Li, Y.; Yang, M.; Tan, Y.; Zhou, L.; Lin, S.; Wen, H. Efficient human activity recognition solving the confusing activities via deep ensemble learning. *IEEE Access* **2019**, *7*, 75490–75499. [CrossRef]
90. Jegham, I.; Khalifa, A.B.; Alouani, I.; Mahjoub, M.A. Vision-based human action recognition: An overview and real world challenges. *Forensic Sci. Int. Digit. Investig.* **2020**, *32*, 200901. [CrossRef]
91. Pareek, P.; Thakkar, A. A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications. *Artif. Intell. Rev.* **2021**, *54*, 2259–2322. [CrossRef]
92. Liu, H.; Chen, S.; Kubota, N. Intelligent Video Systems and Analytics: A Survey. *IEEE Trans. Ind. Inform.* **2013**, *9*, 1222–1233. [CrossRef]
93. Mathur, G.; Bundele, M. Research on Intelligent Video Surveillance techniques for suspicious activity detection critical review. In Proceedings of the 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, India, 23–25 December 2016. [CrossRef]
94. Hou, L.; Liu, Q.; Chen, Z.; Xu, J. Human Detection in Intelligent Video Surveillance: A Review. *J. Adv. Comput. Intell. Intell. Inform.* **2018**, *22*, 1056–1064. [CrossRef]
95. Chaaraoui, A.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]
96. Meinel, L.; Findeisen, M.; Hes, M.; Apitzsch, A.; Hirtz, G. Automated real-time surveillance for ambient assisted living using an omnidirectional camera. In Proceedings of the 2014 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 10–13 January 2014. [CrossRef]
97. Pal, S.; Abhayaratne, C. Video-based Activity Level Recognition for Assisted Living Using Motion Features. In Proceedings of the 9th International Conference on Distributed Smart Cameras, ICDSC '15, Seville, Spain, 8–11 September 2015; ACM: New York, NY, USA, 2015; pp. 62–67. [CrossRef]
98. Rafferty, J.; Nugent, C.D.; Liu, J.; Chen, L. From Activity Recognition to Intention Recognition for Assisted Living Within Smart Homes. *IEEE Trans. Hum.-Mach. Syst.* **2017**, *47*, 368–379. [CrossRef]

99. Koppula, H.S.; Saxena, A. Anticipating Human Activities Using Object Affordances for Reactive Robotic Response. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 14–29. [CrossRef] [PubMed]
100. Ramirez-Amaro, K.; Beetz, M.; Cheng, G. Transferring skills to humanoid robots by extracting semantic representations from observations of human activities. *Artif. Intell.* **2017**, *247*, 95–118. [CrossRef]
101. Rezazadegan, F.; Shirazi, S.; Upcrofit, B.; Milford, M. Action recognition: From static datasets to moving robots. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017. [CrossRef]
102. Tripathi, R.K.; Jalal, A.S.; Agrawal, S.C. Suspicious human activity recognition: A review. *Artif. Intell. Rev.* **2018**, *50*, 283–339. [CrossRef]
103. Reinsel, D.; Gantz, J.; Rydning, J. *Data Age 2025, The Digitization of the World. From Edge to Core*; Technical Report; IDC: Needham, MA, USA, 2018.
104. Bąk, A.; Kulbacki, M.; Segen, J.; Świątkowski, D.; Wereszczyński, K. Recent Developments on 2D Pose Estimation From Monocular Images. In *Intelligent Information and Database Systems*; Nguyen, N.T., Trawiński, B., Fujita, H., Hong, T.P., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; pp. 437–446. [CrossRef]
105. Zheng, C.; Wu, W.; Yang, T.; Zhu, S.; Chen, C.; Liu, R.; Shen, J.; Kehtarnavaz, N.; Shah, M. Deep Learning-Based Human Pose Estimation: A Survey. *arXiv* **2020**, arXiv:2012.13392
106. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv* **2018**, arXiv:1812.08008.
107. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. RMPE: Regional Multi-person Pose Estimation. *arXiv* **2017**, arXiv:1612.00137.
108. Xiu, Y.; Li, J.; Wang, H.; Fang, Y.; Lu, C. Pose Flow: Efficient Online Pose Tracking. *arXiv* **2018**, arXiv:1802.00977.
109. Abdulla, W. Mask R-CNN for Object Detection and Instance Segmentation on Keras and TensorFlow. 2017. Available online: https://github.com/matterport/Mask_RCNN (accessed on 3 June 2022).
110. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
111. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
112. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703. [CrossRef]
113. Pishchulin, L.; Insafutdinov, E.; Tang, S.; Andres, B.; Andriluka, M.; Gehler, P.V.; Schiele, B. Deepcut: Joint subset partition and labeling for multi person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4929–4937. [CrossRef]
114. Toshev, A.; Szegedy, C. Deeppose: Human pose estimation via deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660. [CrossRef]
115. Güler, R.A.; Neverova, N.; Kokkinos, I. Densepose: Dense human pose estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7297–7306. [CrossRef]
116. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.G.; Lee, J.; et al. Mediapipe: A framework for building perception pipelines. *arXiv* **2019**, arXiv:1906.08172.
117. MediaPipe. MediaPipe. 2021. Available online: https://google.github.io/mediapipe/ (accessed on 24 April 2022).
118. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
119. Rahman, M. *Beginning Microsoft Kinect for Windows SDK 2.0: Motion and Depth Sensing for Natural User Interfaces*; Apress: New York, NY, USA, 2017.
120. wrnch Inc. wrnchAI. 2023. Available online: https://wrnch.ai/ (accessed on 24 April 2022).
121. Kendall, A.; Grimes, M.; Cipolla, R. Posenet: A convolutional network for real-time 6-dof camera relocalization. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2938–2946. [CrossRef]
122. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018. [CrossRef]
123. 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/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1227–1236. [CrossRef]
124. Jo, B.; Kim, S. Comparative Analysis of OpenPose, PoseNet, and MoveNet Models for Pose Estimation in Mobile Devices. *Traitement du Signal* **2022**, *39*, 119–124. [CrossRef]
125. Zhao, R.; Wang, K.; Su, H.; Ji, Q. Bayesian graph convolution LSTM for skeleton based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6882–6892. [CrossRef]
126. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1110–1118. [CrossRef]

127. Gong, W.; Zhang, X.; Gonzàlez, J.; Sobral, A.; Bouwmans, T.; Tu, C.; Zahzah, E.H. Human pose estimation from monocular images: A comprehensive survey. *Sensors* **2016**, *16*, 1966. [CrossRef] [PubMed]
128. Sargano, A.; Angelov, P.; Habib, Z. A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. *Appl. Sci.* **2017**, *7*, 110. [CrossRef]
129. Dang, Q.; Yin, J.; Wang, B.; Zheng, W. Deep learning based 2d human pose estimation: A survey. *Tsinghua Sci. Technol.* **2019**, *24*, 663–676. [CrossRef]
130. Munea, T.L.; Jembre, Y.Z.; Weldegebriel, H.T.; Chen, L.; Huang, C.; Yang, C. The progress of human pose estimation: A survey and taxonomy of models applied in 2D human pose estimation. *IEEE Access* **2020**, *8*, 133330–133348. [CrossRef]
131. Gupta, P.; Thatipelli, A.; Aggarwal, A.; Maheshwari, S.; Trivedi, N.; Das, S.; Sarvadevabhatla, R.K. Quo vadis, skeleton action recognition? *Int. J. Comput. Vis.* **2021**, *129*, 2097–2112. [CrossRef]
132. Staniszewski, M.; Kloszczyk, M.; Segen, J.; Wereszczyński, K.; Drabik, A.; Kulbacki, M. Recent Developments in Tracking Objects in a Video Sequence. In *Intelligent Information and Database Systems*; Nguyen, N.T., Trawiński, B., Fujita, H., Hong, T.P., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; pp. 427–436. [CrossRef]
133. Alzughaibi, A.; Chaczko, Z. Efficient Human Motion Detection Feature Set by Using HOG-LPQ Technique. In Proceedings of the 2nd International Congress of Technology, Management and Social Sciences-16 (ICTMS-16), Toronto, ON, Canada, 25–26 June 2016.
134. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition—CVPR 2005, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [CrossRef]
135. Comaniciu, D.; Ramesh, V.; Meer, P. Kernel-based object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 564–577. [CrossRef]
136. Grabner, H.; Grabner, M.; Bischof, H. Real-Time Tracking via On-line Boosting. In Proceedings of the British Machine Vision Conference 2006, Edinburgh, UK, 4–7 September 2006; pp. 47–56. [CrossRef]
137. Avidan, S. Support Vector Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 1064–1072. [CrossRef]
138. Babenko, B.; Yang, M.; Belongie, S.J. Visual tracking with online Multiple Instance Learning. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 983–990. [CrossRef]
139. Jepson, A.D.; Fleet, D.J.; El-Maraghi, T.F. Robust Online Appearance Models for Visual Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 1296–1311. [CrossRef]
140. Santner, J.; Leistner, C.; Saffari, A.; Pock, T.; Bischof, H. PROST: Parallel robust online simple tracking. In Proceedings of the2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 723–730. [CrossRef]
141. Li, X.; Hu, W.; Shen, C.; Zhang, Z.; Dick, A.R.; van den Hengel, A. A Survey of Appearance Models in Visual Object Tracking. *arXiv* **2013**, arXiv:1303.4803. https://doi.org/10.1145/2508037.2508039.
142. Smeulders, A.W.M.; Chu, D.M.; Cucchiara, R.; Calderara, S.; Dehghan, A.; Shah, M. Visual Tracking: An Experimental Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1442–1468. [CrossRef] [PubMed]
143. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.A.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [CrossRef] [PubMed]
144. Ristani, E.; Tomasi, C. Tracking Multiple People Online and in Real Time. In *Computer Vision—ACCV 2014*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2015; Volume 9007, pp. 444–459. [CrossRef]
145. Zamir, A.R.; Dehghan, A.; Shah, M. GMCP-Tracker: Global Multi-object Tracking Using Generalized Minimum Clique Graphs. In *Computer Vision—ECCV 2012*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7573, pp. 343–356. [CrossRef]
146. Dehghan, A.; Assari, S.M.; Shah, M. GMMCP tracker: Globally optimal Generalized Maximum Multi Clique problem for multiple object tracking. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4091–4099. [CrossRef]
147. Ross, G.T.; Soland, R.M. A branch and bound algorithm for the generalized assignment problem. *Math. Program.* **1975**, *8*, 91–103. [CrossRef]
148. Ayazoglu, M.; Sznaier, M.; Camps, O.I. Fast algorithms for structured robust principal component analysis. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1704–1711. [CrossRef]
149. Park, H.; Zhang, L.; Rosen, J.B. Low Rank Approximation of a Hankel Matrix by Structured Total Least Norm. *BIT Numer. Math.* **1999**, *39*, 757–779. [CrossRef]
150. Milan, A.; Leal-Taixé, L.; Schindler, K.; Reid, I.D. Joint tracking and segmentation of multiple targets. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5397–5406. [CrossRef]
151. Poiesi, F.; Cavallaro, A. Tracking Multiple High-Density Homogeneous Targets. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *25*, 623–637. [CrossRef]
152. Bae, S.H.; Yoon, K. Robust Online Multi-object Tracking Based on Tracklet Confidence and Online Discriminative Appearance Learning. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1218–1225. [CrossRef]

153. Kim, T.; Stenger, B.; Kittler, J.; Cipolla, R. Incremental Linear Discriminant Analysis Using Sufficient Spanning Sets and Its Applications. *Int. J. Comput. Vis.* **2011**, *91*, 216–232. [CrossRef]
154. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Accurate Scale Estimation for Robust Visual Tracking. In Proceedings of the British Machine Vision Conference 2014, Nottingham, UK, 1–5 September 2014. [CrossRef]
155. Bolme, D.S.; Beveridge, J.R.; Draper, B.A.; Lui, Y.M. Visual object tracking using adaptive correlation filters. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2544–2550. [CrossRef]
156. Hare, S.; Saffari, A.; Torr, P.H.S. Struck: Structured output tracking with kernels. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 263–270. [CrossRef]
157. Jia, X.; Lu, H.; Yang, M. Visual tracking via adaptive structural local sparse appearance model. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1822–1829. [CrossRef]
158. Zhong, W.; Lu, H.; Yang, M. Robust object tracking via sparsity-based collaborative model. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1838–1845. [CrossRef]
159. Zhang, K.; Zhang, L.; Liu, Q.; Zhang, D.; Yang, M. Fast Visual Tracking via Dense Spatio-temporal Context Learning. In *Computer Vision—ECCV 2014*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8693, pp. 127–141. [CrossRef]
160. Henriques, J.F.; Caseiro, R.; Martins, P.; Batista, J. High-Speed Tracking with Kernelized Correlation Filters. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 583–596. [CrossRef]
161. Gudyś, A.; Rosner, J.; Segen, J.; Wojciechowski, K.; Kulbacki, M. Tracking People in Video Sequences by Clustering Feature Motion Paths. In *Computer Vision and Graphics: International Conference, ICCVG 2014, Warsaw, Poland, 15–17 September 2014*; Springer: Cham, Switzerland, 2014; pp. 236–245. [CrossRef]
162. Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; Ling, H. Lasot: A high-quality benchmark for large-scale single object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5374–5383.
163. Huang, L.; Zhao, X.; Huang, K. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1562–1577. [CrossRef]
164. Leal-Taixé, L.; Milan, A.; Reid, I.; Roth, S.; Schindler, K. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. *arXiv* **2015**, arXiv:1504.01942.
165. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A Benchmark for Multi-Object Tracking. *arXiv* **2016**, arXiv:1603.00831.
166. Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; Leal-Taixé, L. MOT20: A benchmark for multi object tracking in crowded scenes. *arXiv* **2020**, arXiv:2003.09003.
167. Voigtlaender, P.; Krause, M.; Osep, A.; Luiten, J.; Sekar, B.B.G.; Geiger, A.; Leibe, B. MOTS: Multi-Object Tracking and Segmentation. *arXiv* **2019**, arXiv:1902.03604.
168. Dave, A.; Khurana, T.; Tokmakov, P.; Schmid, C.; Ramanan, D. TAO: A Large-Scale Benchmark for Tracking Any Object. In *Computer Vision—ECCV 2020*; Springer: Cham, Switzerland, 2020. [CrossRef]
169. Sun, P.; Cao, J.; Jiang, Y.; Zhang, R.; Xie, E.; Yuan, Z.; Wang, C.; Luo, P. Transtrack: Multiple object tracking with transformer. *arXiv* **2020**, arXiv:2012.15460.
170. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8126–8135.
171. Ma, F.; Shou, M.Z.; Zhu, L.; Fan, H.; Xu, Y.; Yang, Y.; Yan, Z. Unified Transformer Tracker for Object Tracking. *arXiv* **2022**, arXiv:2203.15175.
172. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent models of visual attention. *arXiv* **2014**, arXiv:1406.6247.
173. Bian, T.; Hua, Y.; Song, T.; Xue, Z.; Ma, R.; Robertson, N.; Guan, H. VTT: Long-term Visual Tracking with Transformers. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 9585–9592.
174. Yan, B.; Peng, H.; Fu, J.; Wang, D.; Lu, H. Learning Spatio-Temporal Transformer for Visual Tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 10448–10457.
175. Dunnhofer, M.; Simonato, K.; Micheloni, C. Combining complementary trackers for enhanced long-term visual object tracking. *Image Vis. Comput.* **2022**, *122*, 104448. [CrossRef]
176. Marvasti-Zadeh, S.M.; Cheng, L.; Ghanei-Yakhdan, H.; Kasaei, S. Deep learning for visual tracking: A comprehensive survey. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 3943–3968. [CrossRef]
177. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011. [CrossRef]
178. 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.
179. Marszałek, M.; Laptev, I.; Schmid, C. Actions in context. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 2929–2936. [CrossRef]

180. Kulbacki, M.; Segen, J.; Wereszczyński, K.; Gudyś, A. VMASS: Massive Dataset of Multi-camera Video for Learning, Classification and Recognition of Human Actions. In *Intelligent Information and Database Systems: 6th Asian Conference, ACIIDS 2014, Bangkok, Thailand, 7–9 April 2014, Proceedings, Part II*; Springer: Cham, Switzerland, 2014; pp. 565–574. [CrossRef]
181. Li, W.; Wong, Y.; Liu, A.A.; Li, Y.; Su, Y.T.; Kankanhalli, M. Multi-camera action dataset for cross-camera action recognition benchmarking. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 187–196.
182. Aktı, Ş.; Tataroğlu, G.A.; Ekenel, H.K. Vision-based fight detection from surveillance cameras. In Proceedings of the 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), Istanbul, Turkey, 6–9 November 2019; pp. 1–6. [CrossRef]
183. Cheng, M.; Cai, K.; Li, M. RWF-2000: An Open Large Scale Video Database for Violence Detection. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4183–4190. [CrossRef]
184. Wojciechowski, S.; Kulbacki, M.; Segen, J.; Wyciślok, R.; Bąk, A.; Wereszczyński, K.; Wojciechowski, K. Selected Space-Time Based Methods for Action Recognition. In *Intelligent Information and Database Systems*; Nguyen, N.T., Trawiński, B., Fujita, H., Hong, T.P., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; pp. 417–426. [CrossRef]
185. Ballan, L.; Bertini, M.; Bimbo, A.D.; Serra, G. Video Event Classification Using Bag of Words and String Kernels. In *Image Analysis and Processing–ICIAP 2009*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 170–178. [CrossRef]
186. Bilen, H.; Fernando, B.; Gavves, E.; Vedaldi, A.; Gould, S. Dynamic Image Networks for Action Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 3034–3042. [CrossRef]
187. Laptev, I.; Lindeberg, T. Space-time Interest Points. In Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV 2003), Nice, France, 14–17 October 2003; pp. 432–439. [CrossRef]
188. Shi, F.; Laganière, R.; Petriu, E.M. Gradient Boundary Histograms for Action Recognition. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015, Waikoloa, HI, USA, 5–9 January 2015; pp. 1107–1114. [CrossRef]
189. Wang, H.; Ullah, M.M.; Kläser, A.; Laptev, I.; Schmid, C. Evaluation of Local Spatio-temporal Features for Action Recognition. In Proceedings of the British Machine Vision Conference, BMVC 2009, London, UK, 7–10 September 2009; pp. 1–11. [CrossRef]
190. Zhu, Q.; Yeh, M.C.; Cheng, K.T.; Avidan, S. Fast human detection using a cascade of histograms of oriented gradients. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1491–1498. [CrossRef]
191. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.L. Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vis.* **2013**, *103*, 60–79. [CrossRef]
192. Peng, X.; Wang, L.; Wang, X.; Qiao, Y. Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice. *Comput. Vis. Image Underst.* **2016**, *150*, 109–125
193. Oneata, D.; Verbeek, J.J.; Schmid, C. Action and Event Recognition with Fisher Vectors on a Compact Feature Set. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, 1–8 December 2013; pp. 1817–1824. [CrossRef]
194. Wang, H.; Kläser, A.; Schmid, C.; Liu, C. Action recognition by dense trajectories. In Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3169–3176. [CrossRef]
195. Wang, H.; Schmid, C. Action Recognition with Improved Trajectories. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, 1–8 December 2013; pp. 3551–3558. [CrossRef]
196. Wang, H.; Oneata, D.; Verbeek, J.J.; Schmid, C. A robust and efficient video representation for action recognition. *Int. J. Comput. Vis.* **2016**, *119*, 219–238.
197. Lan, Z.; Lin, M.; Li, X.; Hauptmann, A.G.; Raj, B. Beyond Gaussian Pyramid: Multi-skip Feature Stacking for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; pp. 204–212. [CrossRef]
198. Kantorov, V.; Laptev, I. Efficient Feature Extraction, Encoding, and Classification for Action Recognition. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, 23–28 June 2014; pp. 2593–2600. [CrossRef]
199. Seo, J.; Son, J.; Kim, H.; Neve, W.D.; Ro, Y.M. Efficient and effective human action recognition in video through motion boundary description with a compact set of trajectories. In Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2015, Ljubljana, Slovenia, 4–8 May 2015; pp. 1–6. [CrossRef]
200. Peng, X.; Zou, C.; Qiao, Y.; Peng, Q. Action Recognition with Stacked Fisher Vectors. In Proceedings of the Computer Vision—ECCV 2014—13th European Conference, Part V, Zurich, Switzerland, 6–12 September 2014; pp. 581–595. [CrossRef]
201. Uijlings, J.R.R.; Duta, I.C.; Sangineto, E.; Sebe, N. Video classification with Densely extracted HOG/HOF/MBH features: An evaluation of the accuracy/computational efficiency trade-off. *Int. J. Multimed. Inf. Retr.* **2015**, *4*, 33–44. [CrossRef]
202. Jain, M.; van Gemert, J.C.; Snoek, C.G.M. What do 15, 000 object categories tell us about classifying and localizing actions? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015; pp. 46–55. [CrossRef]

203. 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. [CrossRef]
204. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.
205. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis. (IJCV)* **2015**, *115*, 211–252. [CrossRef]
206. Wang, X.; Gao, L.; Wang, P.; Sun, X.; Liu, X. Two-Stream 3-D convNet Fusion for Action Recognition in Videos with Arbitrary Size and Length. *IEEE Trans. Multimed.* **2018**, *20*, 634–644. [CrossRef]
207. Shi, Y.; Tian, Y.; Wang, Y.; Huang, T. Sequential Deep Trajectory Descriptor for Action Recognition with Three-Stream CNN. *IEEE Trans. Multimedia* **2017**, *19*, 1510–1520. [CrossRef]
208. Wang, L.; Ge, L.; Li, R.; Fang, Y. Three-stream CNNs for action recognition. *Pattern Recognit. Lett.* **2017**, *92*, 33–40. [CrossRef]
209. 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. [CrossRef]
210. 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. [CrossRef]
211. Yang, H.; Yuan, C.; Li, B.; Du, Y.; Xing, J.; Hu, W.; Maybank, S. Asymmetric 3D Convolutional Neural Networks for action recognition. *Pattern Recognit.* **2019**, *85*, 1–12. [CrossRef]
212. 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, CVPR 2015, Boston, MA, USA, 7–12 June 2015; pp. 4305–4314. [CrossRef]
213. Yang, X.; Molchanov, P.; Kautz, J. Multilayer and multimodal fusion of deep neural networks for video classification. In Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 978–987. [CrossRef]
214. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
215. Girdhar, R.; Ramanan, D.; Gupta, A.; Sivic, J.; Russell, B. ActionVLAD: Learning spatio-temporal aggregation for action classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; Volume 2017, pp. 3165–3174. [CrossRef]
216. Xu, Y.; Han, Y.; Hong, R.; Tian, Q. Sequential Video VLAD: Training the Aggregation Locally and Temporally. *IEEE Trans. Image Process.* **2018**, *27*, 4933–4944. [CrossRef]
217. Tu, Z.; Li, H.; Zhang, D.; Dauwels, J.; Li, B.; Yuan, J. Action-Stage Emphasized Spatiotemporal VLAD for Video Action Recognition. *IEEE Trans. Image Process.* **2019**, *28*, 2799–2812. [CrossRef]
218. Binte Naeem, H.; Murtaza, F.; Yousaf, M.; Velastin, S. T-VLAD: Temporal vector of locally aggregated descriptor for multiview human action recognition. *Pattern Recognit. Lett.* **2021**, *148*, 22–28. [CrossRef]
219. Zhao, S.; Liu, Y.; Han, Y.; Hong, R.; Hu, Q.; Tian, Q. Pooling the Convolutional Layers in Deep ConvNets for Video Action Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 1839–1849. [CrossRef]
220. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. *Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.)* **2016**, *9912 LNCS*, 20–36. [CrossRef]
221. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal Segment Networks for Action Recognition in Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2740–2755. [CrossRef] [PubMed]
222. Sun, S.; Kuang, Z.; Sheng, L.; Ouyang, W.; Zhang, W. Optical Flow Guided Feature: A Fast and Robust Motion Representation for Video Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1390–1399. [CrossRef]
223. Bilen, H.; Fernando, B.; Gavves, E.; Vedaldi, A. Action Recognition with Dynamic Image Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2799–2813. [CrossRef]
224. Wang, P.; Li, W.; Li, C.; Hou, Y. Action recognition based on joint trajectory maps with convolutional neural networks. *Knowl.-Based Syst.* **2018**, *158*, 43–53. [CrossRef]
225. Wang, P.; Li, W.; Gao, Z.; Zhang, Y.; Tang, C.; Ogunbona, P. Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; Volume 2017, pp. 416–425. [CrossRef]
226. Yin, C.; Zhu, Y.; Fei, J.; He, X. A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks. *IEEE Access* **2017**, *5*, 21954–21961. [CrossRef]
227. Sun, L.; Jia, K.; Chen, K.; Yeung, D.; Shi, B.; Savarese, S. Lattice Long Short-Term Memory for Human Action Recognition. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; Volume 2017, pp. 2166–2175. [CrossRef]

228. Li, Z.; Gavrilyuk, K.; Gavves, E.; Jain, M.; Snoek, C. VideoLSTM convolves, attends and flows for action recognition. *Comput. Vis. Image Underst.* **2018**, *166*, 41–50. [CrossRef]
229. 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. [CrossRef]
230. Guo, M.H.; Xu, T.X.; Liu, J.J.; Liu, Z.N.; Jiang, P.T.; Mu, T.J.; Zhang, S.H.; Martin, R.R.; Cheng, M.M.; Hu, S.M. Attention mechanisms in computer vision: A survey. *Comput. Visual Media* **2022**, *8*, 331–368. [CrossRef]
231. Long, X.; Gan, C.; De Melo, G.; Wu, J.; Liu, X.; Wen, S. Attention Clusters: Purely Attention Based Local Feature Integration for Video Classification. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7834–7843. [CrossRef]
232. Dai, C.; Liu, X.; Lai, J. Human action recognition using two-stream attention based LSTM networks. *Appl. Soft Comput. J.* **2020**, *86*, 105820. [CrossRef]
233. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 41. [CrossRef]
234. Selva, J.; Johansen, A.S.; Escalera, S.; Nasrollahi, K.; Moeslund, T.B.; Clapés, A. Video Transformers: A Survey. *arXiv* **2022**, arXiv:2201.05991.
235. Girdhar, R.; Carreira, J.; Doersch, C.; Zisserman, A. Video action transformer network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 244–253.
236. Neimark, D.; Bar, O.; Zohar, M.; Asselmann, D. Video transformer network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3163–3172.
237. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The long-document transformer. *arXiv* **2020**, arXiv:2004.05150.
238. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding. *arXiv* **2021**, arXiv:2102.05095.
239. Plizzari, C.; Cannici, M.; Matteucci, M. Spatial temporal transformer network for skeleton-based action recognition. In Proceedings of the International Conference on Pattern Recognition, Shanghai, China, 15–17 October 2021; pp. 694–701.
240. Serrano, S.; Smith, N.A. Is attention interpretable? *arXiv* **2019**, arXiv:1906.03731.
241. Jain, S.; Wallace, B.C. Attention is not explanation. *arXiv* **2019**, arXiv:1902.10186.
242. Wiegreffe, S.; Pinter, Y. Attention is not not explanation. *arXiv* **2019**, arXiv:1908.04626.
243. Chefer, H.; Gur, S.; Wolf, L. Transformer interpretability beyond attention visualization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 782–791.
244. Chaquet, J.M.; Carmona, E.J.; Fernández-Caballero, A. A survey of video datasets for human action and activity recognition. *Comput. Vis. Image Underst.* **2013**, *117*, 633–659. [CrossRef]
245. Olugbade, T.; Bieńkiewicz, M.; Barbareschi, G.; D'Amato, V.; Oneto, L.; Camurri, A.; Holloway, C.; Björkman, M.; Keller, P.; Clayton, M.; et al. Human Movement Datasets: An Interdisciplinary Scoping Review. *ACM Comput. Surv.* **2022**, *55*, 1–29. [CrossRef]
246. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing Human Actions: A Local SVM Approach. In Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3, Cambridge, UK, 26 August 2004; IEEE Computer Society: Washington, DC, USA, 2004; pp. 32–36. [CrossRef]
247. Wu, P.; Liu, J.; Shi, Y.; Sun, Y.; Shao, F.; Wu, Z.; Yang, Z. Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020. [CrossRef]
248. Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; Basri, R. Actions as Space-Time Shapes. In Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV 2005), Beijing, China, 17–20 October 2005; pp. 1395–1402. [CrossRef]
249. Rodriguez, M.D.; Ahmed, J.; Shah, M. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AK, USA, 24–26 June 2008. [CrossRef]
250. Liu, J.; Luo, J.; Shah, M. Recognizing realistic actions from videos. In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009; pp. 1996–2003. [CrossRef]
251. Sultani, W.; Chen, C.; Shah, M. Real-world anomaly detection in surveillance videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6479–6488. [CrossRef]
252. Niebles, J.C.; Chen, C.; Li, F. Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. In Proceedings of the 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; Proceedings, Part II; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6312, pp. 392–405. [CrossRef]
253. Reddy, K.K.; Shah, M. Recognizing 50 human action categories of web videos. *Mach. Vis. Appl.* **2013**, *24*, 971–981. [CrossRef]
254. Yeung, S.; Russakovsky, O.; Jin, N.; Andriluka, M.; Mori, G.; Fei-Fei, L. Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos. *Int. J. Comput. Vis.* **2017**, *126*, 375–389. [CrossRef]
255. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.Y.; Kot, A.C. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2684–2701. [CrossRef] [PubMed]
256. Carreira, J.; Noland, E.; Hillier, C.; Zisserman, A. A short note on the kinetics-700 human action dataset. *arXiv* **2019**, arXiv:1907.06987.

257. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
258. Yao, G.; Lei, T.; Zhong, J. A review of Convolutional-Neural-Network-based action recognition. *Pattern Recognit. Lett.* **2019**, *118*, 14–22. [CrossRef]
259. Liu, Y.; Zhang, Y.; Wang, Y.; Hou, F.; Yuan, J.; Tian, J.; Zhang, Y.; Shi, Z.; Fan, J.; He, Z. A Survey of Visual Transformers. *arXiv* **2021**, arXiv:2111.06091.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.