Mathematical and Computational Applications

MDPI

Review

An Overview of the Vision-Based Human Action Recognition Field

Fernando Camarena *, Miguel Gonzalez-Mendoza *, Leonardo Chang and Ricardo Cuevas-Ascencio

Tecnologico de Monterrey, School of Engineering and Science, Av. Eugenio Garza Sada 2501 Sur, Tecnológico, 64700 Monterrey, Mexico

* Correspondence: fernando@camarenat.com (F.C.); mgonza@tec.mx (M.G.-M.)

Abstract: Artificial intelligence's rapid advancement has enabled various applications, including intelligent video surveillance systems, assisted living, and human–computer interaction. These applications often require one core task: video-based human action recognition. Research in human video-based human action recognition is vast and ongoing, making it difficult to assess the full scope of available methods and current trends. This survey concisely explores the vision-based human action recognition field and defines core concepts, including definitions and explanations of the common challenges and most used datasets. Additionally, we provide in an easy-to-understand manner the literature approaches and their evolution over time, emphasizing intuitive notions. Finally, we explore current research directions and potential future paths. The core goal of this work is to provide future works with a shared understanding of fundamental ideas and clear intuitions about current works and find new research opportunities.

Keywords: video-based human action recognition; action recognition; deep learning methods; handcrafted methods; human action; overview

Citation: Camarena, F.; Gonzalez-Mendoza, M.; Chang, L.; Cuevas-Ascencio, R. An Overview of the Vision-Based Human Action Recognition Field. *Math. Comput. Appl.* **2023**, *28*, 61. https://doi.org/10.3390/mca28020061

Academic Editors: Efrén Mezura-Montes and Nicholas Fantuzzi

Received: 31 January 2023
Revised: 28 March 2023
Accepted: 6 April 2023
Published: 13 April 2023

1. Introduction

Artificial intelligence (AI) redefines our understanding of the world by enabling high-impact applications such as intelligent video surveillance systems [1], self-driving vehicles [2], and assisted living [3]. In addition, AI is revolutionizing areas such as education [4], healthcare [5], abnormal activity recognition [6], sports [7], entertainment [4,8], and human–computer interface systems [9]. These applications frequently rely upon the core task of video-based human action recognition, an active research field to extract meaningful information by detecting and recognizing what a subject is doing in a video [10–12]. Since its critical role in computer vision applications, the action recognition study can lead to innovative solutions that can benefit society in various ways. Nevertheless, it can take time to introduce oneself to the subject thoroughly.

On the one hand, current research points out numerous directions, including effectively combining multi-modal information [13,14], learning without annotated labels [15], training with reduced data points [15,16], and exploring novel architectures [17,18].

On the other hand, recent surveys shifted their focus towards comprehensively analyzing a particular contribution. For example, Ref. [8] categorized standard vision-based human action recognition datasets, whereas Ref. [19] analyzes the classification performance of standard action recognition algorithms. Ref. [20] was one of the first surveys to review deep learning algorithms, providing a comprehensive overview of the datasets employed. Ref. [21] offers a comprehensive taxonomy centered on deep learning methodologies, while Refs. [22,23] concentrates on its applicability. Ref. [24] explores human action recognition from visual and non-visual modalities. Ref. [25] provides proper taxonomy for action transformers according to their architecture, modality, and intended use. Ref. [26] evaluates existing solutions based on the computer vision challenge they solve. Ref. [22]

explores the action recognition field in conjunction with the related tasks of action detection and localization. Finally, Ref. [27] delved into future directions of the field.

Considering the vast expanse of knowledge and numerous potential directions within video-based human action recognition, introducing oneself to the subject requires significant time to develop a comprehensive understanding. As a result, this work strives to offer a comprehensive and intuitive overview of the vision-based human action recognition field:

- In Section 2, we start by defining core concepts, including definitions and explanations of the common challenges and most used datasets that may help future researchers have a shared understanding of the fundamental ideas.
- In Section 3, we break down the literature approaches and their evolution over time, emphasizing the intuitive notions that underpin the approaches' advancements. Therefore, future research may have a clear intuition of what researchers have proposed, and complex concepts make it more accessible to future works.
- In Section 4, we explore current research directions and potential future paths to help future works identify opportunities and boost the process to build further contributions. Finally, we discuss the conclusions in Section 5.

2. Understanding Video-Based Human Action Recognition

The aim of this section is threefold. First, Section 2.1 explains what this work understands as action. Second, we introduce the common challenges of video-based human action recognition in Section 2.2. Third, in Section 2.3, we introduce the commonly used datasets for action recognition. A summary can be found in Figure 1.

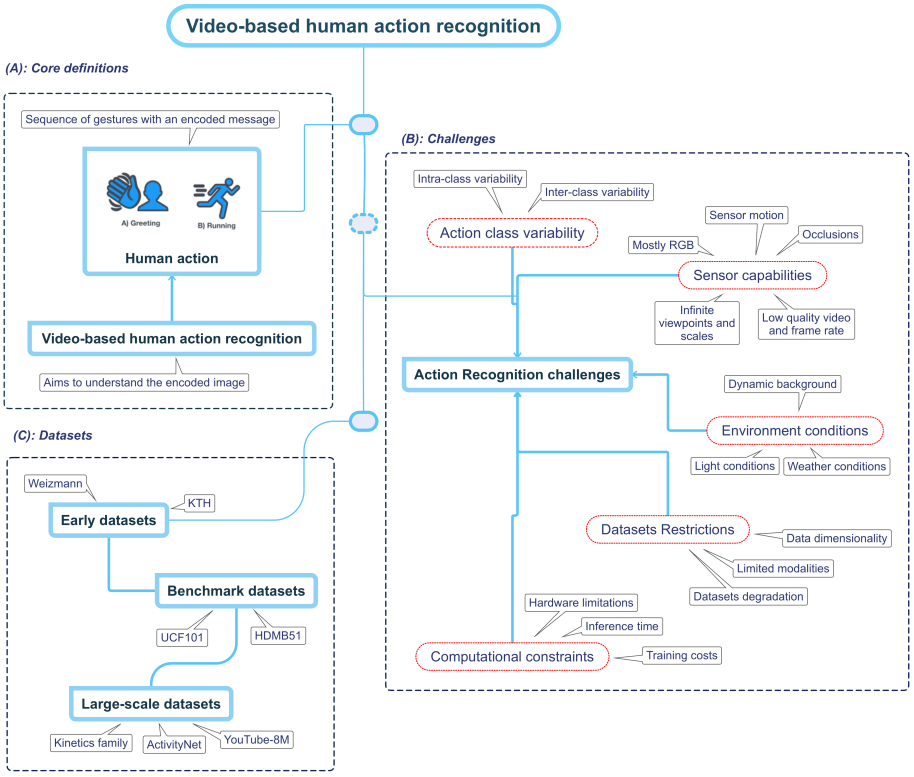

Figure 1. Video-based human action recognition overview. Part (**A**) represents human action; we instinctively associate a sequence of gestures with an action. For example, we might think of the typical hand wave when we think of the action greeting. On the contrary, imagining a person running will create a more dynamic scene with movement centered on the legs. Part (**B**) explains current challenges in the field, and Part (**C**) shows the relevant dataset used.

2.1. What Is an Action?

To understand the idea behind an action, picture the image of a person greeting another. Probably, the mental image constructed involves the well-known waving hand movement. Likewise, if we create a picture of a man running, we may build a more dynamic image by focusing on his legs, as depicted in Figure 1. We unconsciously associate a particular message with a sequence of movements, which we call "an action" [4,28]. In other words, human action is an observable entity that another entity, including a computer, can decode through different sensors. The human action recognition goal is to build approaches to understand the encoded message in the sequence of gestures.

Although it is a natural talent for a person to recognize what others do, it is not an easy assignment for a computer since it faces numerous challenges [20], explained in Section 2.2.

2.2. Challenges Involved in Video-Based Human Action Recognition

While humans have a natural ability to perceive and comprehend actions, computers face various difficulties when recognizing such human actions [26]. We categorize the challenges into five primary categories: action-class variability, sensor capabilities, environment conditions, dataset restrictions, and computational constraints. By understanding these challenges, we may build strategies to overcome them and, consequently, improve the model's performance.

2.2.1. Action Class Variability

Both strong intra- and inter-class variations of an action class represent a challenge for video-based human action recognition [26]. The intra-class variations refer to differences within a particular action class [29]. These variations stem from various factors, such as age, body proportions, execution rate, and anthropometric features of the subject [30]. For example, the running action significantly differs between an older and a younger individual. Additionally, we have repeated some of the actions so many times that we already perform them naturally and unconsciously, making it difficult even for the same person to act precisely the same way twice [26,30]. Finally, cultural contexts can impact how humans act, such as in the case of the greeting action class [31]. Due to the variability, developing a single model that accurately represents all instances of the same class is challenging [26]. Therefore, mitigating intra-class variation is a crucial research area in computer vision to represent all instances of the same class accurately.

Conversely, inter-class variation refers to the dissimilarities between distinct action classes [26], representing a significant challenge because some actions could share major feature vectors [27]. For example, while standing up and sitting down may be perceived as distinct actions, they share the same structure and semantics, making it challenging to differentiate one from another if the model approach does not consider their temporal structure [32]. A similar case is the walking and running actions, which, despite being different, can be seen as variations of the same underlying action. Therefore, to make computer vision applications more accurate and valuable, it is essential to make models that can handle inter-class variations.

2.2.2. Sensor Capabilities

In computer vision, a sensor detects and measures environmental properties such as light, temperature, pressure, and motion to convert them into electrical signals for computer processing [33]. Due to the capture of rich visual information, the RGB camera is the most common sensor used in video-based human action recognition, which senses the light intensity of three color channels (red, green, and blue) [4,33].

Using an RGB camera entails some challenges, including a reduced perspective due to the limited field of view [26], which may cause our target to be partially or not present in the camera field; a partial temporal view of the target subject is known as occlusion [4,26,34] and can be caused either by an object, another subject, the same subject or even the light conditions. Dealing with missing information is difficult because the occlusion may hide

the action's representative features [26]. For example, if a player's legs during a kick are not visible to the camera's field of view throughout a soccer match, it can be challenging to establish if they made contact with the ball.

Furthermore, there is no semantic of how to place the camera sensor, which implies that the target subject can appear in infinite perspectives and scales [35]. On the one hand, some perspectives may not help recognize an action [35,36]; for instance, when a person is reading a book, they will usually hold it in front of them; if the camera viewpoint is the subject's back, it will not perceive the book, and therefore, it will not be able to recognize the action.

On the other hand, our perception of speed is affected by the distance of the object from the camera [37]; even if two objects are moving at the same rate, but one of them is farther away from the camera, our brain will perceive that the farther objects are moving slower, an illusion known as "depth perception distortion" [37]. Earlier, we mentioned that running and walking actions differ in their temporal component, and this scaling effect can affect the accuracy recognition.

Another limitation is the low-video quality that some cameras feature [38], which can lead to a scenario where the target function represents only a few pixels that do not provide enough appearance information or the low camera frame rate does not capture the temporal nature of the action.

Although the camera has fixed placement, it does not imply that it is entirely static [39]; for instance, outdoor cameras are commonly affected by external factors that lead to image motions. Despite this, it may be imperceptible for a human. For a computer, it can be challenging because it may change the appearance features due to the lighting perception or misleading mix of the camera motion with the subject motion.

Another limitation is that the sensors extract only RGB images and, in some cases, audio [24]. Therefore, we are omitting complementary information that can boost the model's capabilities to represent an action class better [24].

2.2.3. Environment Conditions

Environmental conditions can significantly impact the classification accuracy of a model to recognize human actions by affecting the significance of the captured data [4,26]. To illustrate, poor weather conditions such as rain, fog, or snow reduce the target subject's visibility and affect the appearance features extracted. Likewise, in "real" conditions, the target subject will find itself in a scene with multiple objects and entities, which will cause a dynamic, unpredictable, and non-semantic background [26]; the delineation and comprehension of the objective and background can become increasingly complex and challenging when additional factors or variables are presented, which obscure the distinction between the foreground and background. Additionally, environmental conditions can generate image noise that limits representative visual features' extraction and complicates the subject track over time [40].

The environment light is also critical in identifying human actions [26], primarily if the model approach only relies on visual data for feature representation. Lighting conditions can cause subjects to be covered by shadows, resulting in occlusions or areas of high/low contrast, making taking clear, accurate, and visual-consistent pictures of the target subject complex. These circumstances may also result in images differing from those used during model training, confounding the recognition process even further.

2.2.4. Dataset Restrictions

The effectiveness of a machine learning model for recognizing human actions heavily depends on the dataset's quality used in its training phase [41]. The dataset's features, such as the number of samples, diversity, and complexity, are crucial in determining the model's performance. However, using a suitable dataset to boost the model's accuracy takes time and effort [42].

The first approach is constructing the dataset from scratch, ensuring the action samples fit the application's requirements. However, this process can be resource-intensive [42] because most effective machine learning models work under a supervised methodology, and consequently, a labeling process is required [43]. Data labeling [43] involves defining labeling guidelines, class categories, and storage pipelines to further annotate each action sample individually, either manually or by outsourcing to an annotation service to ensure consistent and high-quality data samples.

For some application domains, data acquisition can be challenging due to various factors [44], such as the unique nature of the application, concerns regarding data privacy, or ethical considerations surrounding the use of certain types of data [45]. Consequently, data acquisition can be scarce, insufficient, and unbalanced in the action classes, presenting significant obstacles to developing effective models or conducting meaningful analyses [44].

The second approach involves utilizing well-known datasets with a predefined evaluation protocol, enabling researchers to benchmark their methodology against state-of-the-art techniques. Nevertheless, there are some limitations, including the availability of labeled data; for example, the UCF101 [46] and HMDB51 [47] are one of the most used benchmark datasets [21]. Still, their data dimensionality is insufficient to boost the deep-learning model [48]. Furthermore, current datasets for action recognition face the challenge of adequately representing and labeling every variation of a target action [26], which is nearly impossible due to the immense variability in human movements and environmental factors. This limitation can impact the accuracy and generalizability of action recognition models if the dataset does not represent the same data distribution of the target application [26].

Another main problem with publicly available datasets is their degradation over time [26]; for example, a researcher that aims to use the kinetics dataset [48] must download each video sample from the Internet. However, some download links may no longer work, and specific videos may have been removed or blocked. As a result, accessing the same dataset used in prior research is impossible, leading to inconsistent results [26].

Most of the datasets provide the video along with a textual label tag [13]. Although this is enough to train a model to recognize human action, they have two main limitations. On the one hand, there is no clear intuition that text label tags are the optimal label space for human action recognition [49], particularly in cases where a more nuanced or fine-grained approach to labeling is required or in an application scenario where multi-modal information is available [13]. On the other hand, the exclusive use of RGB information in current datasets overlooks the potential benefits of other input sensors [24], such as depth or infrared sensors, which may provide more detailed and complementary representations of human actions in specific application scenarios.

2.2.5. Computational Constraints

The computational resources required to train and deploy a machine-learning model for video-based human action recognition can pose significant challenges for researchers [13].

Regarding model training, most approaches use a supervised methodology [42] whose performance depends on the data dimensionality, and hyperparameter tuning [50]. Consequently, they involve sophisticated architecture designs [51], leading to over-parameterized models requiring extensive computational resources [51]. The well-known model GPT-3 [52] comprises 175 billion parameters and is estimated to demand 3.14E23 FLOPS of computing power. If a V100 GPU were employed, it would require 355 GPU years to complete and cost roughly USD 4.6 million [52]. Additionally, sometimes researchers work with low-quality data; hence, they need to review and preprocess the dataset before the training process [53], which could be labor-intensive considering the data dimensionality required for model deep learning architectures.

Second, some application domains, such as video surveillance systems, require fast inference responses [27], which can be challenging because the model's complexity can exceed the processing capabilities of the underlying hardware [26,27]. To achieve fast

inference response, the model must analyze and classify video data in a time frame, almost in real-time [26,27].

Other application domains restrict to edge devices that prioritize small factors, portability, and convenience instead of processing power [54]. Some devices cannot perform high-end operations such as 3D convolutions [54].

2.3. Datasets for Video-Based Human Action Recognition

As the field of video-based human action recognition continues to grow, researchers increasingly rely on datasets to benchmark their proposed approaches [26], accelerate model development, and mitigate some of the challenges described in Section 2.2.

Finding a dataset that comprehensively covers all possible matches is nearly impossible. Consequently, the researchers must ensure the feature's dataset covers their application requirements.

Early works in RGB-based approaches used the KTH [55] and Weizmann [56] datasets, commonly known as constrained datasets [22]. Although the video clips provide valuable insights about the action samples, they may only partially represent the complexities and challenges of real-world scenarios since they were artificially recorded in controlled environments [22,32]. In the present state of video-based human action recognition, the KTH [55] and Weizmann [56] datasets no longer represent a challenge because current methods outperform them with nearly 100% of classification accuracy [22].

Conversely, due to the growth of video content on social media, including youtube and movie productions, researchers created datasets with a more comprehensive view of the complexity of human action in natural environments [26]. Two of the most common datasets are the UCF101 [46] and HMDB51 [47] datasets. On the one hand, the UCF101 [46] dataset gathered 13,320 video samples from the youtube dataset and divided it into 101 action categories that contain variations in camera motion, object appearance, and pose, object scale, viewpoint, cluttered background, illumination conditions to mitigate some of the challenges described in Section 2.2.

On the other hand, the HMDB51 [47] dataset consists of 6849 video samples extracted from multiple sources, including movies, Preminger archive, YouTube, and Google Videos, divided into 51 action classes. The HMDB51 [47] provides a comprehensive view of human action in a natural environment with variability in the illumination, subject appearance, and backgrounds.

To date, current approaches have achieved high-accuracy performance on the UCF101 [46] and HMDB51 [47]. Even though they still considered benchmark datasets in action recognition and related tasks [27], including self-supervised action recognition [43], zero-shot action recognition [57], and video generation [58], they have the central problem of data dimensionality since the number of video samples is not enough for deep learning requirements [48].

The Kinetics [48] dataset was introduced to address the limitations of existing action recognition benchmarks. The Kinetics [48] size was several orders of magnitude larger compared to UCF101 [46], and HMDB51 [47], including 400 action classes and 300 thousand video samples. Since its introduction, Kinetics has evolved into a family of datasets, including Kinetics-400 [48], Kinetics-600 [59], and Kinetics-700 [60], each containing at least 400, 600, and 700 video clips, respectively, for their corresponding number of action classes.

Continuing this trend, ActivityNet [61] introduced a large-scale video benchmark with 849 h of untrimmed videos of daily activities divided into 203 activity classes. Additionally, to the label tag, the activity net adds the temporal boundaries of the action sample in the video, which help other related tasks, including temporal action detection and action segmentation.

A prominent dataset is YouTube-8M [62], which contains 350,000 hours of videos with audio divided into 3862 classes. In addition to video-based human action recognition, the dataset can be used for understanding tasks, such as content-based video retrieval

and video summarization. A recent extension, Youtube-8M Segments [63], added 237,000 human-verified segment labels that make the dataset appropriate for temporal localization.

In addition to these RGB-based datasets, the NTU RGB+D [64], Kinectics-Skeleton [65] dataset, and J-HMDB [66] include depth and skeleton information, which can further aid in action recognition with additional information on the spatial and temporal features. On the other hand, MUGEN [67] is a novel dataset with 233,000 unique videos focused on multi-modal research, specifically to understand the relation between audio, video, and text. Finally, the something-something v2 dataset [49] contains 20,847 labeled videos of everyday actions that capture the granularity of video action.

3. The Evolution of Video-Based Human Action Recognition Approaches

This section provides an overview of the evolution of video-based human action recognition. We break down into two parts; first, in Section 3.1, we explain the first family of approaches known as handcrafted approaches. Second, in Section 3.2, we speak about the rise of deep learning approaches.

3.1. Handcrafted Approaches

As described in Figure 2, handcrafted approaches established the foundation for video-based human action recognition, which entails a manual feature engineering process, where human experts manually design features that support a computer to understand.

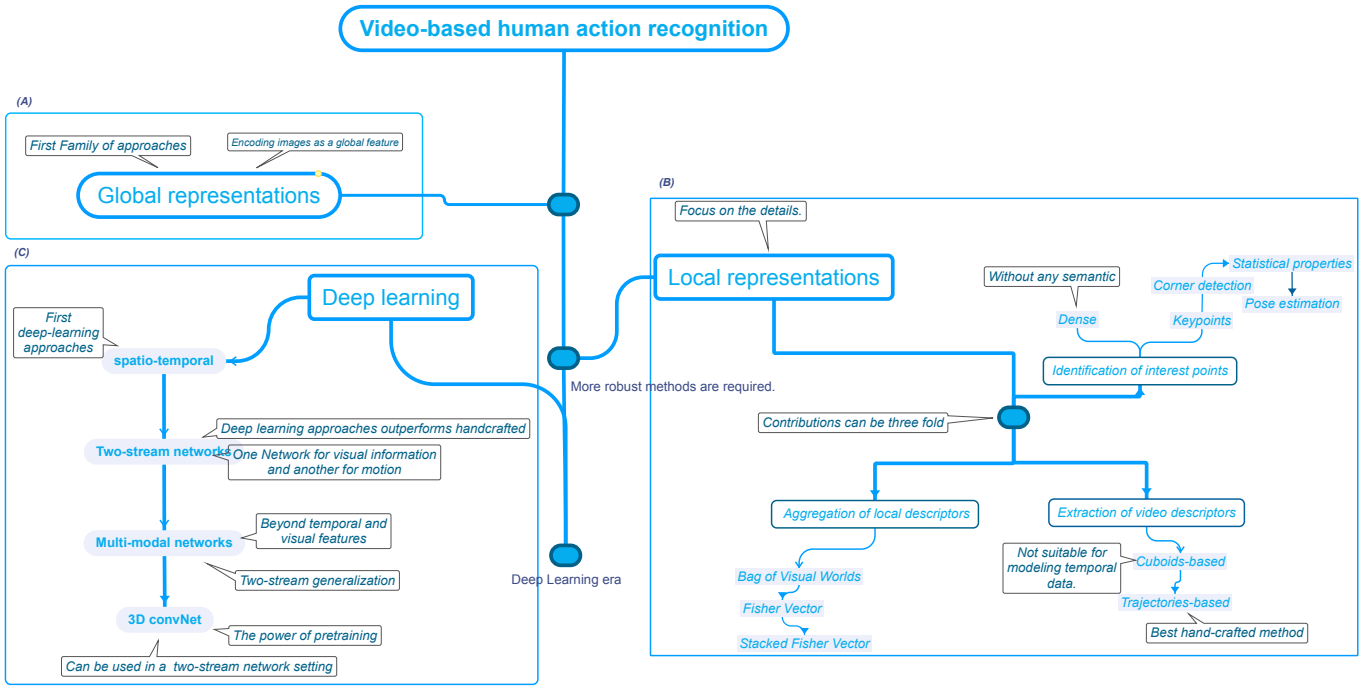

Figure 2. The Evolution of Action Recognition Approaches. The initial attempt at vision-based human action recognition relied on global representations (**A**), which were inferior to local representations (**B**). Lastly, deep learning approaches (**C**) became the most popular, with 3D convolutional neural networks becoming the most advanced because they can learn multiple levels of representations.

Two main components usually form handcrafted approaches. Firstly, feature extraction [4] transforms the input video into a video representation of the action. Secondly, the Action Classification [4] component maps the video representation onto a label tag.

3.1.1. Feature Extraction

Global representations [20] are the first attempt to recognize actions whose intuition is to capture the video input into one global feature. A simple intuition of the effects of

this type of method is our natural ability to recognize human actions only by looking at the subject's silhouette. However, this approach proved inadequate in addressing the numerous challenges posed by videos or images, such as different viewpoints and occlusions. Consequently, global representations could not fully capture the variability of an action. Among the most relevant methods are Motion Energy Image (MEI) [68], Motion History Image (MHI) [69], silhouettes [70], and Spacetime Volume (STV) [71].

The world is full of little details that are difficult to capture using the "big picture". Intuitively, as humans, to discover those little secrets, we need to explore, focus on the details, and zoom in on the regions of interest, which is the idea behind local representations [21,72], as shown in Figure 2B. Local representations seek to extract descriptors from multiple regions of the video to obtain insights into the details. Local approaches break down into a sequence of steps: (a) detection of points of interest, (b) extraction of video descriptors, and (c) aggregations of local descriptors. As a consequence, the researcher's contributions can be three-fold.

As the name suggests, the first step is to detect which regions of the video to analyze. Nevertheless, determining the significance of a region can be a relatively tricky undertaking. Applying edge detection algorithms is one method, such as Space-Time Interest Points (STIPs) [73] and hessian detector [74]. However, its application could lead to noise and lousy performance due to the extraction of edges that belong to something other than the target subject. To assess the regions' relevance and eliminate noisy information, Liu et al. [75] propose using statistical properties as a pruning method.

Camarena et al. [76,77] suggest that pose estimation can be used as the regions of interest, resulting in a method that has a fixed and low number of processing areas, which ensures a consistent frame processing rate. However, the approach is dependent on the subject body's visibility.

Another solution is to apply dense sampling [78], which consists of placing points without semantics. Dense sampling increases the classification accuracy, but it is computationally expensive [76]. In addition, noise injected by other motion sources can affect the classifier's performance [76,77].

Once we have determined which regions to analyze, we must extract the corresponding region description. Visual and motion data are essential for accurately characterizing an action [76]. In this regard, the typical approach combines several descriptors to have a complete perspective of the target action. Regarding the visual information, we have a Histogram Of Oriented Gradients 3D (HOG3D) [79], Speed-Up Robust Features (SURF) [80], 3D SURF [74], and pixel pattern methods [81–83]. On the other hand, descriptors that focus on motion information include Histogram of Oriented Flow (HOF) [84], Motion Boundaries Histogram (MBH) [78], and MPEG flow [85].

Capturing motion information is a complex task; videos are composed of images in which the target person moves or changes location over time [78]. The naive method uses cuboids, which utilize static neighborhood patterns throughout time. However, cuboids are unsuitable for modeling an object's temporal information. Its natural evolution was trajectory-based approaches [78,86,87] that rapidly became one of the most used methods [21,77].

Trajectory-based methods use optical flow algorithms to determine the position of the object of interest in the next frame, which helps to improve the classification performance [21]. Although several efficient optical flow algorithms exist, their application at different points of interest can be computationally expensive [77]. To reduce the computational time, it is essential to know that there are several motion sources besides the subject of interest, including secondary objects, camera motions, and ambient variables. Focusing on the target motion may reduce the amount of computation required. On the one hand, we can use homographies [21] for reducing the motion's camera; on the other hand, pose estimation [77] can be used to remove the optical flow process thoroughly.

Descriptor aggregation is the final stage in which the video descriptor is constructed using the region descriptors acquired from the preceding processes. There are several

methods, including Bag-of-Visual-Words (BoVW) [88], Fisher Vectors [89], Stacked Fisher Vector (SFV) [90], Vector Quantization (VQ) [91], Vector of Locally Aggregated Descriptors (VLAD) [92], Super Vector Encoding (SVC) [93]. Among the handcrafted approaches, it is popularly referred to as FV and SFV, along with dense trajectories achieving the best classification performance [20].

3.1.2. Action Classification

Action classification aims to learn a mapping function to convert a feature vector to a label tag. The literature exposes different approaches, including template-based [68,158,159], generative [160,161], and discriminative models [20,78].

Template-based models are the naive method that compares the feature vector to a set of predefined templates to assign the label tag of the closest instance given a similarity measure. The generative models [160,161] are based on probability and statistics techniques; some representative works include Bayesian Networks and Markov chains.

Discriminative models are one of the most common techniques, including most machine learning methods [20,78]. Due to its performance, handcrafted approaches commonly rely on Support Vector Machines (SVM).

Researchers rely on dimensionality reduction techniques [94] to lower the model's complexity and extract meaningful feature vectors that boost the performance in high-dimensional datasets. Standard techniques include Principal Component Analysis (PCA) [95] and Linear Discriminant Analysis (LDA) [96]. On the one hand, PCA assists in identifying the most representative features, while LDA aids in finding a linear combination of feature vectors that distinguish different action classes.

3.2. Deep Learning Approaches

Due to their strong performance in various computer vision tasks [1–3], Convolutional Neural Networks (CNNs) have become increasingly popular. Hence, its application to vision-based human action recognition appeared inevitable.

Andrej et al. [97] developed one of the first approaches, which involved applying a 2D CNN to each frame and then determining the temporal coherence between the frames. However, unlike other computer vision problems, using a CNN does not outperform handcrafted approaches [27]. The main reason was that human actions are defined by spatial and temporal information, and using a standalone CNN does not fully capture the temporal features [27]. Therefore, subsequent deep learning research for human action recognition has focused on combining temporal and spatial features.

As a common practice, biological processes inspire computer vision and machine learning approaches. For example, as individuals, we use different parts of our brain to process the appearance and motion signals we perceive [98,99]. This understanding can be used for human action recognition, as suggested by [98]. The concept is straightforward. On the one hand, a network extracts spatial characteristics from RGB images. On the other hand, a parallel network extracts motion information from the optical flow output [98]. The network can effectively process visual information by combining spatial and temporal information.

Due to the comparable performance of two-stream networks to trajectory-based methods [27], interest in these approaches grows, leading to novel research challenges such as how to merge the output of motion and appearance features. The most straightforward process, referred to as late fusion [100], is a weighted average of the stream's predictions. More sophisticated solutions considered that interactions between streams should occur as soon as possible and proposed the method of early fusion [100].

Because of the temporal nature of videos, researchers investigated the use of Recurrent Neural Networks (RNN) [101] and Long-Term Short-Term Memory (LSTM) [102,103] as the temporal stream for two-stream approaches. As proven by Ma et al. [104], pre-segmented data are necessary to fully explore the performance of an LSTM in videos thoroughly,

eventually leading to Temporal Segment Networks (TSN), which has become a popular configuration for two-stream networks [27].

A generalization of two-stream networks is multi-stream networks [27], which describe actions using additional modalities such as pose estimation [105], object information [106], audio signals [107], text transcriptions [108], and depth information [109].

One factor that impacts the performance of deep neural networks is the amount of data used to train the model. In principle, the more data we have, the higher our network performance. However, the datasets employed in vision-based human action recognition [46,55,110] do not have the scale that requires a deep learning model [48]. Not disposing of enough data has various implications, one of which is that it is difficult to determine which neural network architecture is optimal. Carreiera et al. [48] introduced the Kinetics dataset as the foundation for re-evaluated state-of-the-art architectures and proposed a novel architecture called Two-Stream Inflated 3D ConvNet (I3D) architecture, based on 2D ConvNet inflation. I3D [48] demonstrates that 3D convolutional networks can be pre-trained, which aids in pushing state-of-the-art action recognition further. Deep learning methods work under a supervised methodology implicating considerable high-quality labels [111]. Nevertheless, data notation is a time-intensive and costly process [111]. Pretraining is a frequent technique to reduce the required processing time and amount of labeled data [111]. Consequently, researchers explored the concept of 2D CNN inflation further [112,113], yielding innovative architectures such as R(2+1)D [114].

Current research in vision-based human action recognition has several directions. First, novel architectures such as visual transformers have been ported to action recognition [114,115]. Second, there is a need for novel training methods such as Self-Supervised Learning (SSL) [43], which is a novel training technique that generates a supervisory signal from unlabeled data, thus eliminating the need for human-annotated labels. Third, few-shot learning action recognition is also being investigated [44].

Most of the architectures described are known as discriminative approaches [116], but there is another family of deep learning methods based on generative techniques [116]. Its core idea is based on the popular phrase "if I cannot create it, then I do not understand it" [117]. Auto-encoders [118], variational autoencoders [119], and Adversarial Networks (GAN) [120] are examples of this approach.

4. Current Research and Future Directions

The video-based human action recognition field is currently undergoing promising research in multiple directions that will shape its future directions.

4.1. New Video-Based Architectures

Since the growing popularity of transformers in natural language processing [25] due to their outstanding capability to process sequential data and superior performance to the well-known convolutional neural networks (CNN) in image-related tasks [25], researchers have been exploring the benefits of visual transformers for human action recognition in video-based applications.

Human action is defined in a visual and temporal space, and understanding the sequential information became crucial to comprehend individual actions and the relationships and dependencies between them [71]. Nevertheless, current methods only focus on the short time frame, which limits the understanding of the impact and consequences of action in the long term, a crucial aspect for the model deployment in real open-world scenarios [22]. Hence, novel architectures should improve an action's visual and temporal information, improving the classification performance.

4.2. Learning Paradigms

A second direction relates to the learning paradigms used to train a model, where supervised learning [42] is the most common; a supervised methodology requires a labeled dataset, meaning every action sample passes through a human-annotated process [50].

Unfortunately, this labeling process is costly and time-consuming, particularly in high-dimensional datasets needed for deep learning approaches [50].

Despite the performance of supervised learning, researchers started to explore new approaches, including semi-supervised learning [121], weakly-supervised learning [122], and Self-Supervised Learning (SSL) [42,43].

Weakly-supervised learning [122] leverages the related information and metadata available on social media, such as hashtags, to approximate the action label tag. On the other hand, the core idea of semi-supervised learning [121] is to extract visual features relying on a small-scale labeled dataset and a large-scale unlabeled dataset. Finally, Self-Supervised Learning (SSL) [42,43] extracts the supervisory signal using the unlabeled data based on the intuition of a child's capability to learn by exploring and interacting with the world.

There are two prominent families of SSL approaches: pretext tasks and contrastive learning. Pretext tasks [123] involve defining an auxiliary function that provides supervised signals without human annotation. For example, a network may be asked if a sequence of video frames is in the correct order. On the other hand, contrastive learning for SSL [162] aims to identify differences between video samples by projecting them onto a shared feature space where clips from the same distribution are clustered together based on a distance metric. Pretext tasks and contrastive learning can be used together, as Pretext Contrastive Learning (PCL) suggests [123]. PCL combines a pretext task function to capture local information and contrastive learning loss functions to gain a global view.

Another challenge related to training a deep learning model is the high correlation between the data used in training and the model performance [50]. On the one hand, it is impractical to construct a novel dataset for each required task. Second, some application domains require highly specific data, such as medical records, making it difficult to recollect a high amount of data [45]. Therefore, few-shot action recognition aims to create models that generalize efficiently in low-data regimes [44].

Its motivation is threefold [124–126]: to enable learning representations for applications where acquiring even unlabeled data is complex, to reduce the high computational demand required for processing large datasets, and to generalize novel action classes not presented in the training dataset. While most few-shot learning research has focused on image tasks [127–129], its extension to video classification is still an open question and remains largely unexplored [125,130].

4.3. Pretraining and Knowledge Transfer

Transferring knowledge from one model to another is a standard technique to reduce computational resources and dependency on labeled data [42]. Traditional methods include transfer learning [131] and fine-tuning [132], which leverage the multi-level representations from deep learning architectures. Transfer learning [131] starts with pre-trained network weights, while fine-tuning [132] adds trainable layers to an existing model. Nevertheless, both transfer methods are model-agnostics meaning that the transfer depends on the model architecture and objective tasks [50]. Novel methods have been explored, including knowledge distillation [50,51,133] that uses a teacher–student framework that enables the transfer even when the new network does not share the same architectural design.

Additionally, to transfer knowledge between different architectural designs, researchers suggest that new methods can lead to transfer learning between different input modalities [134–136]. Sharing knowledge between modalities is challenging, and explorations using disjunct but natural modalities, such as text and visual information, remain a future direction [134]. In addition to using different modalities, novel directions focus on constructing visual models that enable the extraction of visual features that can be representative across multiple video domains in addition to action recognition to construct unified pretraining models [137].

4.4. Video Modalities

Most of the works discussed employ RBG modality; however, incorporating other modalities can benefit various applications scenarios for video-based human action recognition [25]. Modalities can be divided into visual and non-visual [25].

Among visual modalities are RGB [21,134], Skelethon [138,139], depth [24], infrared [140], and thermal [141], each with strengths. For example, the depth [24] modality can extract the objects' shape and structure from the scene. Conversely, infrared [140] modality can capture information in low-light or no-light conditions, and thermal [141] information can detect hidden objects such as humans and temperature monitoring.

One of the most used modalities in video-based human action recognition is skeleton data [138,139,142], which aims to understand human actions using the sequence of the subject skeleton. In contrast to traditional RGB, where Convolutional Neural Networks (CNN) are the standard technique [143], skeleton-based action recognition relies on Graph Neural Networks (GCN) [144,145]. Ref. [143] compares convolutional neural networks to Graph Neural Networks (GCN), showing that proper training techniques, augmentations, and optimizers lead to comparable performance. Ref. [139] presented PYSKL: an open-source toolbox for skeleton-based action recognition that, in addition, to providing CNN and GCN implementations, established a set of good practices to ease the comparison of efficacy and efficiency. Ref. [145] retakes the idea of multiple stream networks and proposes the GCN-Transformer Network (ConGT), which extracts spatial information using the Spatial-Temporal Graph Convolution stream (STG) and temporal information using the Spatial-Temporal Transformer stream (STT).

Regarding non-visual modalities, there are several options, such as audio [146], acceleration [147], radar [148], and WiFi [149], and they are mainly used as complementary data and privacy enhancement. For instance, audio [146] is widely captured along with visual data by video cameras, and it can provide additional and more representative information about some actions, including detecting anomaly events. Radar [148] and WiFi [149] signals 3D-map the environment and understand the object's motion and position in the scene, even in ambient conditions with high levels of occlusions. Finally, the acceleration [147] modality leverages our daily devices' sensors to extract information about motion and body orientation.

In addition to the modalities presented, some may complement our understanding of human action. For example, despite the significant variability of our actions, they have physical limitations, both human and environmental [22]. For this reason, codifying physical properties could lead to a greater understanding of human actions.

4.5. Multi-Modal and Cross-Modal Learning

Speaking about video modalities, our interaction with the world is multi-modal [13], meaning we interact using multiple sensorial inputs.

Therefore, there is no reason that current models use a unique modality. Consequently, leveraging multiple modalities became a new research direction to use the strength of use modality to improve the performance and robustness.

There are two main approaches for using multi-modal learning: multi-modal [43,150] and cross-modal [151]. The foundation of multi-modal learning [43,150] is that diverse modalities can extract different and complementary information that, in conjunction, results in a complete comprehension of the action sample [24]. There are two primary types of approaches to multi-modal learning: fusion [152–154] and co-learning [154]. Fusion [152,153] methods involve merging the classification outputs of models trained separately in different modalities, which can be challenging. In contrast, co-learning [154] aims to use modalities in conjunction with training instead of using them independently, which is more natural to our world perception.

On the other hand, not all modalities are always available simultaneously or are as easy to extract as others [134]. Therefore, cross-modal action recognition aims to transfer knowledge from models trained on different modalities [134], leading to some advantages,

including boosting the performance of a uni-modal model or weaker modality using a stronger modality. Additionally, cross-modal may improve the performance in low-data scenarios [134].

4.6. Explainability

Although deep learning models lead the state-of-the-art in video-based human action recognition, the model outputs are often considered "black boxes" [155], meaning that it is difficult to understand how they make decisions. Some application domains, including video surveillance, imply decisions have real-work consequences; being able to explain the model output is required to be trusted by humans and build transparency in any ethical concern [22]. Explainability in video-based human action recognition is challenging in addition to visual information, and it should include the ability to explain temporal timeframes [22].

It is essential to mention that the current research directions are not standalone paths, and research that combines them is relevant. For example, self-supervised learning is complementary to few-shot learning [156]. Furthermore, knowledge distillation can be used for cross-modal transfer learning [134]. Other relevant directions include studying new data augmentation techniques [157], neural architecture search [4], and efficient network training methods [26]. In addition, constructing a novel dataset that supports previous research directions remains critical to developing novel methods [27].

5. Conclusions

This work provides an overview of the video-based human action recognition field. We started by defining core concepts of the field, including the definition of what an action is and the goal of video-based human action recognition. Then, we described the challenges of action-class variability, sensor capabilities, environment conditions, dataset restrictions, and computational constraints explaining their implications and possible consequences. Finally, we introduced some of the most used datasets in the literature, including traditional RGB-based datasets such as KTH, Weizmann, UCF101, HDMB51, Kinetics, ActivityNet, and YouTube-8M. In addition, we found some datasets that provide additional modalities inputs such as NTU RGB+D, Kinectics-Skeleton, and J-HMDB. The information presented may help future works to have a shared understanding of the fundamental ideas of the field.

Conversely, to provide researchers with a clear intuition of what has been explored and make complex concepts accessible, we explore the approaches proposed in the literature and break down their evolution over time, emphasizing the intuitive notions that underpin the approaches' advancements. The explorations include traditional handcrafted and deep learning approaches. We described local and global feature extraction methods and some standard action classification techniques regarding handcrafted methods. Regarding deep learning methods, we explored traditional methods, two-stream networks, and 3D CNN.

Finally, we explored current research directions and potential paths to help future works identify novel opportunities and boost the process of constructing meaningful contributions. We divided the directions into six blocks: implementation of new architectures, new learning paradigms, new pretraining and transfer methods, exploration of novel modalities, multi-modal and cross-modal, and finally, the model explainability.

Author Contributions: Conceptualization, L.C.; formal analysis, F.C.; investigation, F.C.; methodology, F.C. and M.G.-M.; project administration, M.G.-M.; resources, M.G.-M.; supervision, M.G.-M. and L.C.; validation, M.G.-M.; writing—original draft, F.C. and R.C.-A.; writing—review and editing, F.C. and R.C.-A. All authors have read and agreed to the published version of the manuscript.

Funding: F.C. gratefully acknowledges the scholarship no. 815917 from CONACyT to pursue his postgraduate studies. The scholarship had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors would like to thanks the financial support from Tecnologico de Monterrey through the "Challenge-Based Research Funding Program 2022". Project ID E120-EIC-GI06-B-T3-D.

Conflicts of Interest: The author declares that he has no conflict of interest.

References

1. Elharrouss, O.; Almaadeed, N.; Al-Maadeed, S.; Bouridane, A.; Beghdadi, A. A combined multiple action recognition and summarization for surveillance video sequences. *Appl. Intell.* **2021**, *51*, 690–712. [CrossRef]
2. Badue, C.; Guidolini, R.; Carneiro, R.V.; Azevedo, P.; Cardoso, V.B.; Forechi, A.; Jesus, L.; Berriel, R.; Paixao, T.M.; Mutz, F.; et al. Self-driving cars: A survey. *Expert Syst. Appl.* **2021**, *165*, 113816. [CrossRef]
3. Martinez, M.; Rybok, L.; Stiefelhagen, R. Action recognition in bed using BAMs for assisted living and elderly care. In Proceedings of the 2015 14th IAPR International Conference on Machine Vision Applications (MVA), Tokyo, Japan, 18–22 May 2015; pp. 329–332.
4. 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]
5. Nweke, H.F.; Teh, Y.W.; Mujtaba, G.; Al-Garadi, M.A. Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Inf. Fusion* **2019**, *46*, 147–170. [CrossRef]
6. Nayak, R.; Pati, U.C.; Das, S.K. A comprehensive review on deep learning-based methods for video anomaly detection. *Image Vis. Comput.* **2021**, *106*, 104078. [CrossRef]
7. Ullah, A.; Muhammad, K.; Haq, I.U.; Baik, S.W. Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments. *Future Gener. Comput. Syst.* **2019**, *96*, 386–397. [CrossRef]
8. 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]
9. Rodomagoulakis, I.; Kardaris, N.; Pitsikalis, V.; Mavroudi, E.; Katsamanis, A.; Tsiami, A.; Maragos, P. Multimodal human action recognition in assistive human-robot interaction. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 2702–2706.
10. Meng, Y.; Panda, R.; Lin, C.C.; Sattigeri, P.; Karlinsky, L.; Saenko, K.; Oliva, A.; Feris, R. AdaFuse: Adaptive Temporal Fusion Network for Efficient Action Recognition. *arXiv* **2021**, arXiv:2102.05775.
11. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3316–3333. [CrossRef]
12. Ullah, A.; Muhammad, K.; Hussain, T.; Baik, S.W. Conflux LSTMs network: A novel approach for multi-view action recognition. *Neurocomputing* **2021**, *435*, 321–329. [CrossRef]
13. Alayrac, J.B.; Recasens, A.; Schneider, R.; Arandjelovic, R.; Ramapuram, J.; De Fauw, J.; Smaira, L.; Dieleman, S.; Zisserman, A. Self-Supervised MultiModal Versatile Networks. *NeurIPS* **2020**, *2*, 7.
14. Valverde, F.R.; Hurtado, J.V.; Valada, A. There is More than Meets the Eye: Self-Supervised Multi-Object Detection and Tracking with Sound by Distilling Multimodal Knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11612–11621.
15. Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; Tang, J. Self-supervised learning: Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 857–876. [CrossRef]
16. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.H.; Buchatskaya, E.; Doersch, C.; Pires, B.A.; Guo, Z.D.; Azar, M.G.; et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv* **2020**, arXiv:2006.07733.
17. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *arXiv* **2021**, arXiv:2101.01169.
18. Hafiz, A.M.; Parah, S.A.; Bhat, R.U.A. Attention mechanisms and deep learning for machine vision: A survey of the state of the art. *arXiv* **2021**, arXiv:2106.07550.
19. Cheng, G.; Wan, Y.; Saudagar, A.N.; Namuduri, K.; Buckles, B.P. Advances in human action recognition: A survey. *arXiv* **2015**, arXiv:1501.05964.
20. Zhang, S.; Wei, Z.; Nie, J.; Huang, L.; Wang, S.; Li, Z. A review on human activity recognition using vision-based method. *J. Healthc. Eng.* **2017**, *2017*, 3090343. [CrossRef]
21. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **2017**, *60*, 4–21. [CrossRef]
22. Kong, Y.; Fu, Y. Human action recognition and prediction: A survey. *Int. J. Comput. Vis.* **2022**, *130*, 1366–1401. [CrossRef]
23. Lei, Q.; Du, J.X.; Zhang, H.B.; Ye, S.; Chen, D.S. A survey of vision-based human action evaluation methods. *Sensors* **2019**, *19*, 4129. [CrossRef]
24. Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; Liu, J. Human action recognition from various data modalities: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 3200–3225. [CrossRef]
25. Ulhaq, A.; Akhtar, N.; Pogrebna, G.; Mian, A. Vision Transformers for Action Recognition: A Survey. *arXiv* **2022**, arXiv:2209.05700.
26. 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]
27. Zhu, Y.; Li, X.; Liu, C.; Zolfaghari, M.; Xiong, Y.; Wu, C.; Zhang, Z.; Tighe, J.; Manmatha, R.; Li, M. A comprehensive study of deep video action recognition. *arXiv* **2020**, arXiv:2012.06567.
28. Borges, P.V.K.; Conci, N.; Cavallaro, A. Video-based human behavior understanding: A survey. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *23*, 1993–2008. [CrossRef]

29. Cherla, S.; Kulkarni, K.; Kale, A.; Ramasubramanian, V. Towards fast, view-invariant human action recognition. In Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, Alaska, 23–28 June 2008; pp. 1–8.
30. Stergiou, N.; Decker, L.M. Human movement variability, nonlinear dynamics, and pathology: Is there a connection? *Hum. Mov. Sci.* **2011**, *30*, 869–888. [CrossRef]
31. Matsumoto, D. Cultural similarities and differences in display rules. *Motiv. Emot.* **1990**, *14*, 195–214. [CrossRef]
32. Huang, D.A.; Ramanathan, V.; Mahajan, D.; Torresani, L.; Paluri, M.; Fei-Fei, L.; Niebles, J.C. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7366–7375.
33. Bradski, G.; Kaehler, A. *Learning OpenCV: Computer Vision with the OpenCV Library*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2008.
34. Ramanathan, M.; Yau, W.Y.; Teoh, E.K. Human action recognition with video data: Research and evaluation challenges. *IEEE Trans. -Hum.-Mach. Syst.* **2014**, *44*, 650–663. [CrossRef]
35. Yang, W.; Wang, Y.; Mori, G. Recognizing human actions from still images with latent poses. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2030–2037.
36. Piergiovanni, A.; Ryoo, M.S. Recognizing actions in videos from unseen viewpoints. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 4124–4132.
37. Pfautz, J.D. Distortion of Depth Perception in a Virtual Environment Application. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1996.
38. Demir, U.; Rawat, Y.S.; Shah, M. Tinyvirat: Low-resolution video action recognition. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 7387–7394.
39. Heilbron, F.C.; Thabet, A.; Niebles, J.C.; Ghanem, B. Camera motion and surrounding scene appearance as context for action recognition. In Proceedings of the Computer Vision—ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, 1–5 November 2014; Revised Selected Papers, Part IV 12; Springer: Berlin/Heidelberg, Germany, 2015; pp. 583–597.
40. Kaur, A.; Rao, N.; Joon, T. Literature Review of Action Recognition in the Wild. *arXiv* **2019**, arXiv:1911.12249.
41. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 843–852.
42. Jing, L.; Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4037–4058. [CrossRef]
43. Jaiswal, A.; Babu, A.R.; Zadeh, M.Z.; Banerjee, D.; Makedon, F. A survey on contrastive self-supervised learning. *Technologies* **2020**, *9*, 2. [CrossRef]
44. Kumar Dwivedi, S.; Gupta, V.; Mitra, R.; Ahmed, S.; Jain, A. Protogan: Towards few shot learning for action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
45. Mittelstadt, B.D.; Floridi, L. The ethics of big data: Current and foreseeable issues in biomedical contexts. *Sci. Eng. Ethics* **2016**, *22*, 303–341. [CrossRef] [PubMed]
46. 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.
47. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
48. 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.
49. Goyal, R.; Ebrahimi Kahou, S.; Michalski, V.; Materzynska, J.; Westphal, S.; Kim, H.; Haenel, V.; Fruend, I.; Yianilos, P.; Mueller-Freitag, M.; et al. The something something video database for learning and evaluating visual common sense. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5842–5850.
50. Xu, G.; Liu, Z.; Li, X.; Loy, C.C. Knowledge distillation meets self-supervision. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 588–604.
51. Wang, L.; Yoon, K.J. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3048–3068. [CrossRef] [PubMed]
52. OpenAI's GPT-3 Language Model: A Technical Overview. Available online: https://lambdalabs.com/blog/demystifying-gpt-3 (accessed on 30 January 2023).
53. Kelleher, J.D. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2019.
54. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
55. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition (ICPR), Cambridge, UK, 23–26 August 2004; Volume 3, pp. 32–36.
56. Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time shapes. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Beijing, China, 17–21 October 2005; Volume 2, pp. 1395–1402.
57. Estevam, V.; Pedrini, H.; Menotti, D. Zero-shot action recognition in videos: A survey. *Neurocomputing* **2021**, *439*, 159–175. [CrossRef]

58. Hong, W.; Ding, M.; Zheng, W.; Liu, X.; Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv* **2022**, arXiv:2205.15868.
59. Carreira, J.; Noland, E.; Banki-Horvath, A.; Hillier, C.; Zisserman, A. A short note about kinetics-600. *arXiv* **2018**, arXiv:1808.01340.
60. Carreira, J.; Noland, E.; Hillier, C.; Zisserman, A. A short note on the kinetics-700 human action dataset. *arXiv* **2019**, arXiv:1907.06987.
61. Elmaghraby, S.E. Activity nets: A guided tour through some recent developments. *Eur. J. Oper. Res.* **1995**, *82*, 383–408. [CrossRef]
62. Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; Vijayanarasimhan, S. Youtube-8m: A large-scale video classification benchmark. *arXiv* **2016**, arXiv:1609.08675.
63. Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, A.; Toderici, G.; Varadarajan, B.; Vijayanarasimhan, S. The YouTube-8M Dataset. 2016. Available online: https://research.google.com/youtube8m/ (accessed on 27 March 2023).
64. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
65. Li, J.; Wong, K.S.; Liu, T.T. Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates. In Proceedings of the 27th ACM International Conference on Multimedia, ACM, Nice, France, 21–25 October 2019; pp. 1398–1406.
66. Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; Black, M.J. Towards understanding action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 940–952.
67. Hayes, T.; Zhang, S.; Yin, X.; Pang, G.; Sheng, S.; Yang, H.; Ge, S.; Hu, Q.; Parikh, D. Mugen: A playground for video-audio-text multimodal understanding and generation. In *Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022, Proceedings, Part VIII*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 431–449.
68. Bobick, A.; Davis, J. An appearance-based representation of action. In Proceedings of the 13th International Conference on Pattern Recognition, Vienna, Austria, 25–19 August 1996; Volume 1, pp. 307–312.
69. Huang, C.P.; Hsieh, C.H.; Lai, K.T.; Huang, W.Y. Human action recognition using histogram of oriented gradient of motion history image. In Proceedings of the 2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control, Beijing, China, 21–23 October 2011; pp. 353–356.
70. Li, W.; Zhang, Z.; Liu, Z. Action recognition based on a bag of 3d points. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 9–14.
71. Poppe, R. A survey on vision-based human action recognition. *Image Vis. Comput.* **2010**, *28*, 976–990. [CrossRef]
72. Rodríguez-Moreno, I.; Martínez-Otzeta, J.M.; Sierra, B.; Rodriguez, I.; Jauregi, E. Video activity recognition: State-of-the-art. *Sensors* **2019**, *19*, 3160. [CrossRef] [PubMed]
73. Laptev, I. On space-time interest points. *Int. J. Comput. Vis.* **2005**, *64*, 107–123. [CrossRef]
74. Willems, G.; Tuytelaars, T.; Van Gool, L. An efficient dense and scale-invariant spatio-temporal interest point detector. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 650–663.
75. Liu, J.; Luo, J.; Shah, M. Recognizing realistic actions from videos "in the wild". In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1996–2003.
76. Camarena, F.; Chang, L.; Gonzalez-Mendoza, M. Improving the Dense Trajectories Approach Towards Efficient Recognition of Simple Human Activities. In Proceedings of the 2019 7th International Workshop on Biometrics and Forensics (IWBF), Cancun, Mexico, 2–3 May 2019; pp. 1–6.
77. Camarena, F.; Chang, L.; Gonzalez-Mendoza, M.; Cuevas-Ascencio, R.J. Action recognition by key trajectories. *Pattern Anal. Appl.* **2022**, *25*, 409–423. [CrossRef]
78. Wang, H.; Kläser, A.; Schmid, C.; Liu, C.L. Action recognition by dense trajectories. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 3169–3176.
79. Klaser, A.; Marszałek, M.; Schmid, C. A Spatio-Temporal Descriptor Based on 3d-Gradients. Available online: https://class.inrialpes.fr/pub/klaser-bmvc08.pdf (accessed on 30 January 2023).
80. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]
81. Zhao, G.; Pietikainen, M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 915–928. [CrossRef]
82. Norouznezhad, E.; Harandi, M.T.; Bigdeli, A.; Baktash, M.; Postula, A.; Lovell, B.C. Directional space-time oriented gradients for 3d visual pattern analysis. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 736–749.
83. Tuzel, O.; Porikli, F.; Meer, P. Region covariance: A fast descriptor for detection and classification. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 589–600.
84. Dalal, N.; Triggs, B.; Schmid, C. Human detection using oriented histograms of flow and appearance. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 428–441.
85. Kantorov, V.; Laptev, I. Efficient feature extraction, encoding and classification for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2593–2600.
86. Messing, R.; Pal, C.; Kautz, H. Activity recognition using the velocity histories of tracked keypoints. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 104–111.

87. Matikainen, P.; Hebert, M.; Sukthankar, R. Trajectons: Action recognition through the motion analysis of tracked features. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, Japan, 29 September–2 October 2009; pp. 514–521.
88. Chang, L.; Pérez-Suárez, A.; Hernández-Palancar, J.; Arias-Estrada, M.; Sucar, L.E. Improving visual vocabularies: A more discriminative, representative and compact bag of visual words. *Informatica* **2017**, *41*, 333–347.
89. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In Proceedings of the European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 143–156.
90. Peng, X.; Zou, C.; Qiao, Y.; Peng, Q. Action recognition with stacked fisher vectors. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 581–595.
91. Sivic, J.; Zisserman, A. Video Google: A text retrieval approach to object matching in videos. In Proceedings of the Computer Vision, IEEE International Conference, Nice, France, 13–16 October 2003; p. 1470.
92. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
93. Zhou, X.; Yu, K.; Zhang, T.; Huang, T.S. Image classification using super-vector coding of local image descriptors. In Proceedings of the European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 141–154.
94. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [CrossRef]
95. Pearson, K. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
96. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]
97. 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.
98. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *arXiv* **2014**, arXiv:1406.2199.
99. Goodale, M.A.; Milner, A.D. Separate visual pathways for perception and action. *Trends Neurosci.* **1992**, *15*, 20–25. [CrossRef] [PubMed]
100. Ye, H.; Wu, Z.; Zhao, R.W.; Wang, X.; Jiang, Y.G.; Xue, X. Evaluating two-stream CNN for video classification. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, 23–26 June 2015; pp. 435–442.
101. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
102. Dai, C.; Liu, X.; Lai, J. Human action recognition using two-stream attention based LSTM networks. *Appl. Soft Comput.* **2020**, *86*, 105820. [CrossRef]
103. Gammulle, H.; Denman, S.; Sridharan, S.; Fookes, C. Two stream lstm: A deep fusion framework for human action recognition. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 177–186.
104. Ma, C.Y.; Chen, M.H.; Kira, Z.; AlRegib, G. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Process. Image Commun.* **2019**, *71*, 76–87. [CrossRef]
105. Choutas, V.; Weinzaepfel, P.; Revaud, J.; Schmid, C. Potion: Pose motion representation for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7024–7033.
106. Ikizler-Cinbis, N.; Sclaroff, S. Object, scene and actions: Combining multiple features for human action recognition. In Proceedings of the European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 494–507.
107. He, D.; Li, F.; Zhao, Q.; Long, X.; Fu, Y.; Wen, S. Exploiting spatial-temporal modelling and multi-modal fusion for human action recognition. *arXiv* **2018**, arXiv:1806.10319.
108. Hsiao, J.; Li, Y.; Ho, C. Language-guided Multi-Modal Fusion for Video Action Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3158–3162.
109. Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time human action recognition based on depth motion maps. *J. Real-Time Image Process.* **2016**, *12*, 155–163. [CrossRef]
110. Wang, L.; Qiao, Y.; Tang, X. Action recognition with trajectory-pooled deep-convolutional descriptors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4305–4314.
111. Tao, L.; Wang, X.; Yamasaki, T. Pretext-Contrastive Learning: Toward Good Practices in Self-supervised Video Representation Leaning. *arXiv* **2020**, arXiv:2010.15464.
112. Wang, X.; Miao, Z.; Zhang, R.; Hao, S. I3d-lstm: A new model for human action recognition. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Kazimierz Dolny, Poland, 21–23 November 2019; IOP Publishing: Bristol, UK, 2019; Volume 569, p. 032035.
113. Liu, G.; Zhang, C.; Xu, Q.; Cheng, R.; Song, Y.; Yuan, X.; Sun, J. I3D-Shufflenet Based Human Action Recognition. *Algorithms* **2020**, *13*, 301. [CrossRef]

114. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.
115. Chen, J.; Ho, C.M. MM-ViT: Multi-modal video transformer for compressed video action recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 1910–1921.
116. Beddiar, D.R.; Nini, B.; Sabokrou, M.; Hadid, A. Vision-based human activity recognition: A survey. *Multimed. Tools Appl.* **2020**, *79*, 30509–30555. [CrossRef]
117. Gleick, J. *Genius: The Life and Science of Richard Feynman*; Vintage: New York, NY, USA, 1993.
118. Xing, L.; Qin-kun, X. Human action recognition using auto-encode and pnn neural network. *Softw. Guide* **2018**, *1*, 1608-01529.
119. Mishra, A.; Pandey, A.; Murthy, H.A. Zero-shot learning for action recognition using synthesized features. *Neurocomputing* **2020**, *390*, 117–130. [CrossRef]
120. Ahsan, U.; Sun, C.; Essa, I. Discrimnet: Semi-supervised action recognition from videos using generative adversarial networks. *arXiv* **2018**, arXiv:1801.07230.
121. Zhu, X.; Goldberg, A.B. Introduction to semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **2009**, *3*, 1–130.
122. Zhou, Z.H. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **2018**, *5*, 44–53. [CrossRef]
123. Tao, L.; Wang, X.; Yamasaki, T. Selfsupervised video representation using pretext-contrastive learning. *arXiv* **2020**, arXiv:2010.15464.
124. Xing, J.; Wang, M.; Mu, B.; Liu, Y. Revisiting the Spatial and Temporal Modeling for Few-shot Action Recognition. *arXiv* **2023**, arXiv:2301.07944.
125. Gowda, S.N.; Sevilla-Lara, L.; Kim, K.; Keller, F.; Rohrbach, M. A new split for evaluating true zero-shot action recognition. In Proceedings of the Pattern Recognition: 43rd DAGM German Conference, DAGM GCPR 2021, Bonn, Germany, 28 September–1 October 2021; Springer: Berlin/Heidelberg, Germany, 2022; pp. 191–205.
126. Li, S.; Liu, H.; Qian, R.; Li, Y.; See, J.; Fei, M.; Yu, X.; Lin, W. TA2N: Two-stage action alignment network for few-shot action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 1404–1411.
127. Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; Darrell, T. Few-shot object detection via feature reweighting. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8420–8429.
128. Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv. (CSUR)* **2020**, *53*, 1–34. [CrossRef]
129. Dong, N.; Xing, E.P. Few-Shot Semantic Segmentation with Prototype Learning. In Proceedings of the BMVC, Newcastle, UK, 3–6 September 2018; Volume 3.
130. Cao, K.; Ji, J.; Cao, Z.; Chang, C.Y.; Niebles, J.C. Few-shot video classification via temporal alignment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10618–10627.
131. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 1–40. [CrossRef]
132. Ribani, R.; Marengoni, M. A survey of transfer learning for convolutional neural networks. In Proceedings of the 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), Rio de Janeiro, Brazil, 28–31 October 2019; pp. 47–57.
133. Gou, J.; Yu, B.; Maybank, S.J.; Tao, D. Knowledge Distillation: A Survey. *arXiv* **2020**, arXiv:2006.05525.
134. Zhen, L.; Hu, P.; Peng, X.; Goh, R.S.M.; Zhou, J.T. Deep multimodal transfer learning for cross-modal retrieval. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *33*, 798–810. [CrossRef] [PubMed]
135. Rajan, V.; Brutti, A.; Cavallaro, A. Cross-modal knowledge transfer via inter-modal translation and alignment for affect recognition. *arXiv* **2021**, arXiv:2108.00809.
136. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (PMLR), Virtual, 18–24 July 2021; pp. 8748–8763.
137. Zhu, W.; Ma, X.; Liu, Z.; Liu, L.; Wu, W.; Wang, Y. MotionBERT: Unified Pretraining for Human Motion Analysis. *arXiv* **2022**, arXiv:2210.06551.
138. Duan, H.; Wang, J.; Chen, K.; Lin, D. Pyskl: Towards good practices for skeleton action recognition. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 7351–7354.
139. Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; Dai, B. Revisiting skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2969–2978.
140. Akula, A.; Shah, A.K.; Ghosh, R. Deep learning approach for human action recognition in infrared images. *Cogn. Syst. Res.* **2018**, *50*, 146–154. [CrossRef]
141. Batchuluun, G.; Nguyen, D.T.; Pham, T.D.; Park, C.; Park, K.R. Action recognition from thermal videos. *IEEE Access* **2019**, *7*, 103893–103917. [CrossRef]
142. Wang, X.; Zhang, S.; Qi, G.; Wu, Y.; Wu, Y.; Tang, S.; Zhang, J.; Zhang, Y. View-Invariant Skeleton-based Action Recognition via Global-Local Contrastive Learning. *arXiv* **2021**, arXiv:2103.06751.
143. Ali, A.; Pinyoanuntapong, E.; Wang, P.; Dorodchi, M. Skeleton-based Human Action Recognition via Convolutional Neural Networks (CNN). *arXiv* **2023**, arXiv:2301.13360.

144. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
145. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13359–13368.
146. Gao, R.; Oh, T.H.; Grauman, K.; Torresani, L. Listen to look: Action recognition by previewing audio. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10457–10467.
147. Haider, F.; Salim, F.A.; Postma, D.B.; Van Delden, R.; Reidsma, D.; van Beijnum, B.J.; Luz, S. A super-bagging method for volleyball action recognition using wearable sensors. *Multimodal Technol. Interact.* **2020**, *4*, 33. [CrossRef]
148. Yang, S.; Le Kernec, J.; Fioranelli, F. *Action Recognition Using Indoor Radar Systems*; IET Human Motion Analysis for Healthcare Applications: London, UK, 2019.
149. Guo, J.; Shi, M.; Zhu, X.; Huang, W.; He, Y.; Zhang, W.; Tang, Z. Improving human action recognition by jointly exploiting video and WiFi clues. *Neurocomputing* **2021**, *458*, 14–23. [CrossRef]
150. Schiappa, M.C.; Rawat, Y.S.; Shah, M. Self-supervised learning for videos: A survey. *ACM Comput. Surv.* **2022**. [CrossRef]
151. Thoker, F.M.; Gall, J. Cross-modal knowledge distillation for action recognition. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 6–10.
152. Rani, S.S.; Naidu, G.A.; Shree, V.U. Kinematic joint descriptor and depth motion descriptor with convolutional neural networks for human action recognition. *Mater. Today Proc.* **2021**, *37*, 3164–3173. [CrossRef]
153. Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; Huang, J. Deep multimodal fusion by channel exchanging. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 4835–4845.
154. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [CrossRef]
155. Bai, X.; Wang, X.; Liu, X.; Liu, Q.; Song, J.; Sebe, N.; Kim, B. Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognit.* **2021**, *120*, 108102. [CrossRef]
156. Rajasegaran, J.; Khan, S.; Hayat, M.; Khan, F.S.; Shah, M. Self-supervised knowledge distillation for few-shot learning. *arXiv* **2020**, arXiv:2006.09785.
157. Nida, N.; Yousaf, M.H.; Irtaza, A.; Velastin, S.A. Video augmentation technique for human action recognition using genetic algorithm. *ETRI J.* **2022**, *44*, 327–338. [CrossRef]
158. Rabiner, L.R.; Juang, B.-H. *Fundamentals of Speech Recognition*; PTR Prentice Hall: Englewood Cliffs, NJ, USA, 1993; Volume 14.
159. 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]
160. Natarajan, P.; Nevatia, R. Online, real-time tracking and recognition of human actions. In Proceedings of the 2008 IEEE Workshop on Motion and Video Computing, Copper Mountain, CO, USA, 8–9 January 2008.
161. Oliver, N.M.; Rosario, B.; Pentland, A.P. A Bayesian computer vision system for modeling human interactions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 831–843. [CrossRef]
162. Le-Khac, P.H.; Healy, G.; Smeaton, A.F. Contrastive representation learning: A framework and review. *IEEE Access* **2020**, *8*, 193907–193934. [CrossRef]