

Review

A Survey of Vision-Based Human Action Evaluation Methods

Qing Lei ^{1,2,*}, Ji-Xiang Du ^{1,2,*}, Hong-Bo Zhang ^{1,2}, Shuang Ye ^{1,2} and Duan-Sheng Chen ^{1,2}

¹ Department of Computer Science and Technology, Huaqiao University, Xiamen 361000, China; zhanghongbo@hqu.edu.cn (H.-B.Z.); shuangy_amoy@163.com (S.Y.); dschen@hqu.edu.cn (D.-S.C.)

² Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Huaqiao University, Xiamen 361000, China

* Correspondence: leiqing@hqu.edu.cn (Q.L.); jxdu@hqu.edu.cn (J.-X.D.)

Received: 16 August 2019; Accepted: 18 September 2019; Published: 24 September 2019



Abstract: The fields of human activity analysis have recently begun to diversify. Many researchers have taken much interest in developing action recognition or action prediction methods. The research on human action evaluation differs by aiming to design computation models and evaluation approaches for automatically assessing the quality of human actions. This line of study has become popular because of its explosively emerging real-world applications, such as physical rehabilitation, assistive living for elderly people, skill training on self-learning platforms, and sports activity scoring. This paper presents a comprehensive survey of approaches and techniques in action evaluation research, including motion detection and preprocessing using skeleton data, handcrafted feature representation methods, and deep learning-based feature representation methods. The benchmark datasets from this research field and some evaluation criteria employed to validate the algorithms' performance are introduced. Finally, the authors present several promising future directions for further studies.

Keywords: human action evaluation; action quality assessment; feature learning; hand-crafted features; deep learning features; action evaluation dataset

1. Introduction

In recent years, human action evaluation has emerged as an important problem in a variety of computer vision applications, which range from sports training [1–5] to healthcare and physical rehabilitation [6–9], interactive entertainment [10–12], and video understanding [13–15]. In contrast to the aims of traditional human action recognition to infer the class label from predefined action categories (action classification [16,17]), to locate the starting and end positions of actions (action detection [18,19]), and to predict the future state of actions on the basis of incomplete action observations (action prediction [20,21]), the target of human action evaluation is to make computers automatically quantify how well people perform actions and further provide interpretable feedback for improving the motion of the human body. The field of human action evaluation research has recently begun to diversify because of its explosively emerging real-world applications, including healthcare and physical rehabilitation [22], skill training for expertise learners [23], as well as sports activity scoring [24], as shown in Figure 1.

- (1) Healthcare and rehabilitation. Physical therapy is essential for the recovery training of stroke survivors and sports injury patients. As supervision by a professional provided by a hospital or medical agency is expensive, home training with a vision-based rehabilitation system is preferred since it allows people to practice movement training privately and economically. In

computer-aided therapy or rehabilitation, human action evaluation can be applied to assist patients training at home, guide them to perform actions properly, and prevent them from further injuries.

- (2) Skill training. Human action evaluation plays an important role in assessing the skill level of expert learners on self-learning platforms. For example, simulation-based surgical training platforms have been developed for surgical education. The technique of action quality assessment makes it possible to develop computational approaches that automatically evaluate the surgical students' performance. Accordingly, meaningful feedback information can be provided to individuals and guide them to improve their skill levels. Another application field is the assembly line operations in industrial production. Human action evaluation can help to construct standardized action models related to different operation steps, as well as to evaluate the performance of trained workers. Automatically assessing the action quality of workers can be beneficial by improving working performance, promoting productive efficiency, and, more importantly, discovering dangerous actions before damage occurs.
- (3) Sports activity scoring. In recent years, the video assistant referee (VAR) system has been introduced to some international sports competition events. The VAR serves as a match official who reviews the referee staffing's decisions on the basis of video footage. Thus, decision making can be affected by contacting the referee on the field of play. Human action evaluation makes it possible to analyze the motion quality of athletes, judge the normalization of postures or body movements, and score the action performances automatically. The VAR system facilitates detailed assessments of sports activities and ensures the fairness of competition.

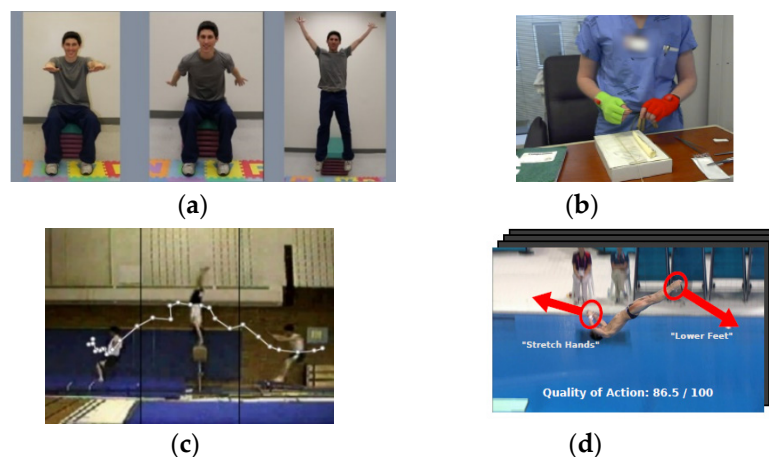


Figure 1. Several application fields of human action evaluation. (a) healthcare and rehabilitation; (b) surgical skill training; (c) sports skill training; (d) sports activity scoring.

Human action evaluation relies on accurate human motion detection and tracking, action segmentation and representation, and efficient evaluation methods for measuring the quality of action performance. Almost all the challenges of human motion analysis also need to be tackled in action evaluation research, including the diversity and complexity of human body movement, action segmentation in long-duration video, and robust feature learning. Some factors increase the difficulties in this field of research, such as intra-class variations in scale, appearance, speed, and style, as well as several environmental factors, including poor illumination, camera motion, and cluttered backgrounds.

In the last decade, quite a few reviews have been published to summarize the advancement of human action recognition. Conversely, a few action evaluation surveys have emerged, except for the work of [25], which is a recent thorough investigation on vision-based action understanding in application to assistive healthcare. The authors reviewed challenges and difficulties in advanced vision-based assistive healthcare and rehabilitation methods, available sensing devices for data

acquisition, as well as some benchmark datasets. However, implementation methods and technical details in this field of research have not been addressed. This paper provides a comprehensive overview of recent human action evaluation studies. The review covers a variety of practical uses not only in the field of healthcare and physical rehabilitation, but also in application to skill training of expertise learners and sports activity scoring.

Most of the reviewed studies have employed three processing stages in their implementation approaches. The first stage is to detect and segment human actions from long-duration video data. The second stage is to extract effective features for evaluating the quality of human actions in video segments, and the third stage is to develop an accurate assessment method for measuring the performance of human actions. The aggregation of segmental predictions to form a final decision on the complete activity is also performed in the final stage. Therefore, this study presents a thorough review of human action evaluation research from the perspective of its major processes and learning methodologies. Accordingly, the reviewed methods are divided into three parts: Motion detection and preprocessing methods, handcrafted feature representation methods, and deep neural network (DNN)-based feature methods for human action evaluation. The classification framework is shown in Figure 2.

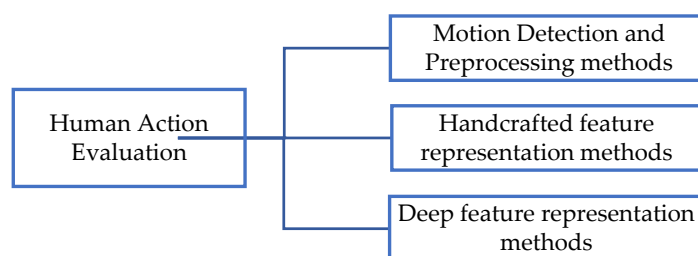


Figure 2. The framework of reviewed action evaluation research works.

Contrary to human action recognition, the research on action evaluation aims not only to recognize the action that is performed but also to provide a quality assessment result and feedback about how the action was performed. It is reasonable to assume that the quality of human action is directly dependent on the dynamic changes in human body movement, which can be represented by the motion trajectory and relative location relationship between joints or body parts, while it is less influenced by environmental aspects, such as cluttered backgrounds or dynamic scenes. Thus, the skeleton data provide significant information for developing reliable quality assessment algorithms for action evaluation. Most related works have employed skeleton data that have been captured or detected from a depth or color camera for action evaluation research. This paper provides a comprehensive overview of recent methods of vision-based skeleton data analysis for action evaluation research, and we compare their performance on benchmark datasets.

The contributions of this paper are fourfold:

- (1) A clear problem definition is given of human action evaluation and clarify the differences among the three tasks of human activity understanding: Action recognition, action prediction, and action evaluation. To the best of the authors' knowledge, it is the first review of action evaluation research that involves the diversity of real-world applications, such as healthcare, physical rehabilitation, skill training, and sports activity scoring.
- (2) A thorough review of human action evaluation methods, including motion detection and preprocessing methods, handcrafted feature representation methods, and deep feature representation methods, are presented in this paper.
- (3) Some benchmark datasets are introduced and compared with the existing research works on these datasets to provide useful information for those interested in this field of research.
- (4) Moreover, some suggestions for future research are presented.

The remainder of this paper is organized as follows. In Section 2, the authors provide the problem definition of human action evaluation and clarify the difference among three tasks in research that aims to understand human activity, namely, action recognition, action prediction, and action evaluation. A thorough review of human action evaluation methods is presented in Section 3, including the skeleton data-based motion detection and preprocessing methods, handcrafted feature representation methods, and deep learning-based feature methods. In Section 4, the benchmark datasets and evaluation criteria for action evaluation are discussed. Finally, our conclusions are drawn in Section 5 and some suggestions are presented for future research.

2. Problem Definition

In human activity analysis, action is usually regarded as human body motion that involves the movements of joints, limbs, or body parts. In an early study on human activity analysis, Aggarwal [26] conceptually categorized human activities into four different levels: Gestures, actions, interactions, and group activities. Comparing the objectives of inference reveals that the research on human activity analysis can generally be divided into three tasks: Action recognition, action prediction, and action quality evaluation. These concepts are introduced below and clarify the differences among them in detail.

Human action recognition (HAR) [27–30] has been put forward to determine what the action is and when it is occurring. Two classic problems of action recognition have been extensively investigated in the past several decades: Action classification [31–36] and action detection [37–41]. The former aims to infer the action label for an input image or video, and the latter aims to detect the starting and ending times of all action occurrences from the video. Since particular types of actions can be recognized depending solely on limb morphology and the surrounding environment, still image or keyframe discovery-based human action recognition approaches have also been extensively researched in the last decade.

Human action prediction (HAP) infers the action label of a video using a partially observed or incomplete action occurrence. The researchers in this field have commonly regarded it as a before-the-fact computer vision task [27,42,43]. It is important to predict abnormal behaviors early in public areas [44,45] or prevent dangerous driving behavior for the sake of traffic safety [46,47]. Human action prediction can be beneficial to some real-time monitoring applications, such as assistive living for elderly people, visual surveillance in public places, and autonomous vehicles.

Human action evaluation (HAE) [1–10,48–50] aims to design computation models for representing the dynamic changing process of human movement and to develop evaluation techniques to measure the completion quality of human action. The target of action evaluation not only requires the recognition of the action performed but also more significantly, needs to provide a quality assessment of how the action was performed. Further, semantic feedback is presented understandably to facilitate the performers' improvement of their deficiencies (see Figure 3).

The major differences between action recognition, action prediction, and action evaluation can be summarized in the following four points:

- (1) From the perspective of the data to be processed, the temporal series of RGB video or depth sequences are a common data type in both action prediction and action evaluation. Still images are barely used for these two tasks. However, action recognition can usually be directly inferred from a still image or key frame of a video sequence. Furthermore, action prediction infers the class label with the condition that only parts of the action occurrence have been observed (the first part or middle part of the complete video). In contrast, accurate action evaluation estimation can only be achieved if complete human body movements have been observed and considered.
- (2) Of the different targets among the three research tasks, human action prediction focuses on discovering abnormal events by forecasting unseen human actions and distinguishing anomalous actions from normal ones. Action recognition tries to classify a given still image or video sequence into a predefined action category or to find the start frame and the end frame of an action

- occurrence from a video sequence. Action evaluation has been presented to make a computer automatically give a good or bad assessment for action quality and provide interpretable feedback.
- (3) Both human motion and environmental backgrounds provide important spatial–temporal information for action recognition and prediction. In contrast, the quality of action performance is directly determined by the dynamic changes in the human body or limbs, while it is less affected by environmental aspects. In particular, in the task of recognizing human–object interactions, such as people using telephones or playing instruments, the interacting objects provide an important clue for inferring the action category or intention. However, in a diving scoring application, the quality score of a diving performance involves the whole movement of the body and limbs of an athlete between taking off and entering the water. It matters whether the athletes’ feet are together and whether their toes are pointed straight throughout the whole diving process. The splash upon entering the water has a relatively minor impact only on the final stage assessment.
 - (4) For facilitating action recognition research, large-scale datasets have been collected and provided to researchers in this field, such as ImageNet, UCF-101, HMDB51, and Sports-1M. Efficient action recognition methods, especially recent deep feature approaches, including 3D convolutional networks, recurrent neural networks, and graph convolutional neural networks, have been trained from these large-scale annotated datasets. On the contrary, only a few small-scale datasets oriented to some specific applications, such as physical therapy, healthcare, sports activity scoring, and skill training, have been available for the research on human action evaluation. Small-scale annotated training samples, limited action categories, a stationary camera position, and the same backgrounds are common shortcomings among the published action evaluation datasets.

The following section presents a comprehensive review of the current human action evaluation research, including the detection and preprocessing methods of skeleton data (including detection, normalization, and alignment), handcrafted feature representation methods, and deep feature representation methods.

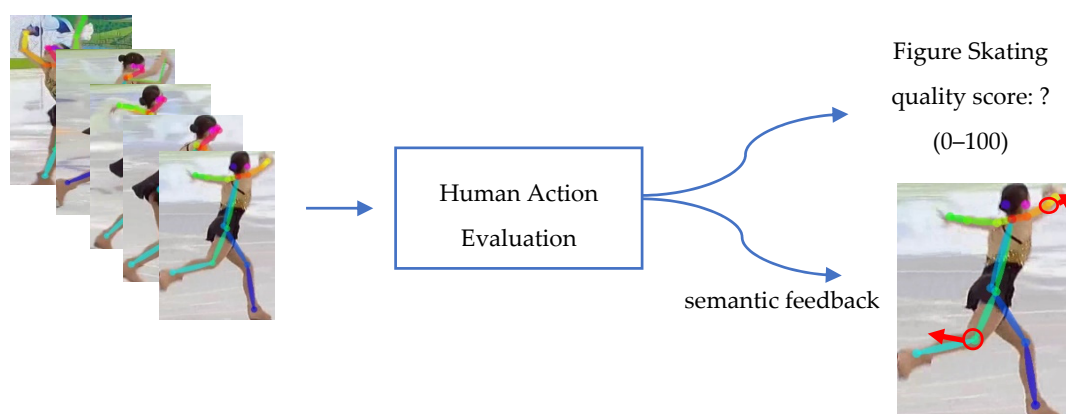


Figure 3. The task of human action evaluation (not only identifying action labels but also assessing the quality score and providing semantic feedback).

3. Overview of Human Action Evaluation Methods

Human action recognition has been investigated by researchers for a long time. Over the past few decades, computer vision-based action recognition techniques and methodologies have grown rapidly. An increasing number of reliable and efficient approaches have advanced quickly with the help of large-scale data and high-performance computation boosting. However, the research on action evaluation has achieved more interest recently because of its wide-ranging applications, such as assistive living for elderly people, medical therapy for patients, and skill training for expertise learners. Gordon [24] first put forward the problem of automated video assessment for human performance. The author proposed using computers that automatically analyze a performance recorded in video data.

Ilg et al. [51] applied an earlier attempt to use the traditional feature extraction method to estimate the sports skill levels of individuals from 3D trajectory data obtained from a motion capture system.

Human activity analysis in action evaluation research also faces challenges, such as the diversity and complexity of human movements, camera motion, occlusion, and background clutter. Similar key issues, including the motion detection and segmentation, feature extraction, and feature similarity measurement, have to be tackled for action evaluation. In early studies in this field, most research works [1–7] directly introduced and fine-tuned state-of-the-art action recognition approaches to deal with action evaluation problems. They either regarded action evaluation as a classification problem or simply replaced the optimization function of classification with regressions in the final stage without analyzing the essential differences between two tasks. Thus, the effect is far from meeting application requirements. Recently, deep neural network-based learning approaches, such as the 3D convolutional network (C3D) [52,53], long short-term memory network (LSTM) [54,55], and recurrent neural network (RNN) [56], have greatly boosted the action recognition performance on large-scale datasets. Some deep learning-based action evaluation studies have been attempted in the past few years.

It is a natural and reasonable assumption that action quality directly depends on the dynamic changes in human body movements, as discussed in previous sections. Thus, skeleton data-based action quality assessment approaches have dominated in this field of research. In earlier studies, skeleton data could only be captured from a conditioned laboratory with environmental settings, such as a motion capture system (MoP), depth cameras, and stereo cameras. Thus, the evaluation performance was strongly restricted by the limitation of a small-scale training dataset. Now, with the help of start-of-the-art pose estimation techniques and methodologies, skeleton data can also be obtained from RGB images that facilitate the diversity of datasets.

From the perspective of application domains, the motion representation and learning methodologies, the reviewed methods are summarized as presented in Table 1. Further, a detailed introduction about the implementation and technical details of typical methods are presented in the following subsections.

3.1. Detection and Preprocessing of Skeleton Data

3.1.1. Skeleton Data Detection

Skeleton-based representations are suitable for modeling human actions by encoding the relationship between the joints or body parts, as well as the holistic body configuration. In order to develop an accurate action evaluation method, it is preferred that the dynamic changes in joints or body parts for analysis and action representation be used. The first important issue is the acquisition and preprocessing of skeleton data. In previous studies, RGB cameras were mostly used to create computer vision datasets. Skeleton data can be obtained from a particular motion capture system, for which performers have to wear markers near each joint to identify the motion by the positions or angles between the markers. As a result of the advent of depth-sensing cameras, such as the Microsoft Kinect sensor and Asus Xmotion, it is much easier to obtain skeleton data, thereby popularizing skeleton-based human representation. However, depth cameras severely suffer from occlusions, sensing distance, and poor performance in outdoor environments.

With significant progress in recent pose estimation techniques and methodologies, skeleton data can now be estimated from RGB image data, which greatly facilitate the diversity of datasets. Traditional skeletonization models, such as the deformable part model [57] and flexible mixtures of parts model [58], have been replaced by deep neural network-based approaches since the advent of DeepPose. In 2014, Toshev and Szegedy [59] first applied deep neural networks for the precise localization of articulated joints for human pose estimation. They employed a two-stage process: A seven-layer generic convolutional DNN was constructed to regress the location of each body joint in the first stage, and in the second stage, DNN-based regressors were learned to refine the joint predictions by using higher-resolution subimages. DeepPose has achieved the best results, which have been superior

to those of most traditional skeletonization models, and it has become the baseline for subsequent deep learning-based human pose estimation methods, such as OpenPose [60], AlphaPose [61], and DensePose [62].

OpenPose [60], the first real-time multi-person system, was developed by the perceptual computing lab of Carnegie Mellon University. It employs the architecture of two-branch multistage CNNs and uses part affinity fields (PAFs) to encode the location and orientation of limbs. These feature vectors sufficiently represent a global context and help to achieve high-quality greedy parsing results. OpenPose provides the functionality of 2D real-time multi-person keypoint detection (15- or 18- or 25-keypoint body/foot keypoint estimation), as well as the 3D real-time single-person keypoint detection. An example of an 18-keypoint skeleton model of OpenPose is shown in Figure 4a, including the nose, neck, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, right hip, right knee, right ankle, left hip, left knee, left ankle, right eye, left eye, right ear, and left ear. Some detection examples of diving and figure skating using the Olympic Sports Dataset [1] are illustrated in Figure 4b,c.

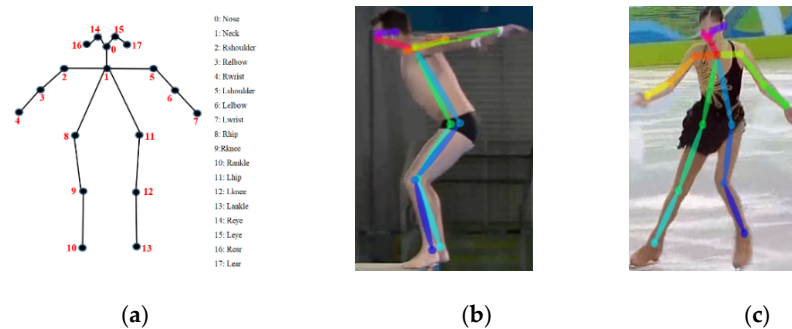


Figure 4. Skeleton model and detection examples of OpenPose [60]. (a) Eighteen-joint skeleton model of OpenPose; (b) example of diving; (c) example of figure skating.

3.1.2. Skeleton Data Preprocessing

The estimated skeleton data can often be noisy in cases of occlusion or cluttered backgrounds in realistic scenes. To obtain robust similarity quantization for fine-grained quality assessment, the preliminary treatment of skeleton data is noise filtering. Subsequently, normalization and alignment [63–67] are performed to deal with intra-class variations in the scale, transition, and rotation, as well as various appearances or motion styles.

In the case of occlusion or self-occlusion, the failure or false detection of the human body leads to the emergence of outliers of the joint's coordinates in skeleton pose estimation that lead to an unreliable representation of human movements. Thus, the noise filtering process has been commonly performed on the original joint coordinates in most research works. The traditional image filtering techniques, such as Laplacian smoothing, Gaussian filtering, discrete cosine transformation (DCT), and discrete Fourier transformation (DFT), have been employed to transform the discrete coordinates of the joint trajectory. Consequently, zero values or sharply changing coordinates can be filtered out, and low-frequency components are preserved for the reliable detection of human body positions.

Since both the height of persons and the photographing distance are quite different, the scale of the human body in different videos can be quite diverse. The original skeleton positions need to be normalized to a prototypical range for comparison. Without scaling, two motion instances with similar action quality may produce very different assessment results. The scale normalization process can be summarized in three steps. First, the middle of the left and right hip in the first skeleton of the sequence is chosen as a reference point. Then, the distance from the neck to that reference point is computed and defined as the normalized length. Finally, the position coordinates of joints are transformed by scaling with the normalized length.

Another intra-class variation is that the spatial and temporal position of the human body can be very different located in various videos. The performer might not be in the same position relative to

the camera, and the action video is usually untrimmed to include the preparing and ending duration of an action instance. Consequently, the skeleton positions need to be aligned in both the spatial and temporal dimensions. Since the position offset relative to the center of the camera will not be the same, the joints' relative positions are computed by subtracting the coordinate of the hip center to obtain a skeleton-centric representation. After that, the rotation transformation is performed on the relative joints' positions. The rotation angle θ is determined by the projection of the vector from the left hip to the right hip onto the x-axis. Then, each joint's coordinate is transformed by rotating θ degrees to remove the impact of view variation. For temporal alignment, the traditional dynamic time warping (DTW) method has been used by some previous research works.

Most published skeleton-based action evaluation studies have performed the normalization and alignment process using the above-introduced methods. Paiement et al. [6] extracted skeleton data from Kinect's depth data and preprocessed the skeleton data with scaling and spatial alignment. They employed diffusion maps, which constitute a graph-based technique with a quasi-isometric mapping of the original higher space to a reduced low-dimensional diffusion space, to reduce the dimensionality and deal with noise and outliers of skeleton data extracted from Kinect's depth data. They validated their method by the assessment of gait on stairs. Pirsiavash et al. [1] employed the flexible mixtures of parts model and dynamic programming algorithm to find the best pose track in an Olympic diving or figure skating video. A discrete cosine transform was performed on the normalized positions of the joint trajectory, and the k lowest frequency components were used as high-level pose features for assessing the quality of human actions in videos. They found that low frequency feature extraction helped remove high-frequency noise due to pose estimation errors. In the work of [22], the authors conducted scaling and spatial alignment to normalize the positions of the joint trajectory, and relative positions and orientations were computed from the raw skeleton 3D time-series data. They also employed discrete cosine transformation to project joint positions and angles into the frequency domain.

3.2. Handcrafted Feature Methods for Human Action Evaluation

The traditional action recognition's two-staged processing inspired a similar framework that comprises action feature representation, and feature assessment learning was commonly used in the early action evaluation research. First, the local or holistic motion features are extracted to represent human actions through some handcraft-designed feature engines. Traditional feature detectors and descriptors, such as spatial-temporal interest points (STIP) [68], histogram of gradient (HOG) [69], histogram of optical flow (HOF) [70], scale-invariant feature transform (SIFT) [71], and motion boundary histogram (MBH) [72], have been commonly employed. Then, some research works formulated action quality assessment into an optimization framework in which action models are trained using the bag of words (BOW) [73] or the hidden Markov model (HMM) [74], and the corresponding evaluation functions are determined to assess the quality score of action features. The interpretable feedback is further provided to improve motion performance. From the perspective of major application fields, a detailed review of handcrafted feature methods in action evaluation research is presented in this subsection.

3.2.1. Handcrafted Features for Physical Rehabilitation

Of the research on physical therapy or rehabilitation, Chen et al. [75] made an early attempt to develop a computational model for the quantitative quality evaluation of human body movement in a rehabilitation clinic setting. They modeled the evaluation function as a linear combination of 33 normalized kinematic attributes organized into seven categories—four activity level categories and three body-function-level categories—such as temporal profile, targeting, trajectory, velocity, and joint function. The RankSVM algorithm was modified and used to estimate the weights of different movements according to the clinician's comparison of impaired subjects' movement quality in pairs. The authors applied the learned kinematic impairment measure function to assess the reach and grasp

of stroke survivors. The high Pearson correlations between the estimated results and the therapist's decisions were achieved.

Venkataraman et al. [76] proposed the use of the shape distribution of the reconstructed phase space called a dynamical attractor as the dynamical feature for modeling actions. They developed a shape-theoretic method to estimate the movement quality for single-marker home-based stroke rehabilitation. They mapped a one-dimensional time series to the reconstructed phase space and extracted the shape distribution to represent the dynamical shape feature of the mapped m-dimensional attractor. Then, the super vector regression function was trained to compute the quality score from dynamical shape features. The Pearson correlation coefficients achieved were better than those in the work on kinematic analysis in [75].

In the work of [77], Celiktutan et al. presented a graph-based physical exercise action evaluation method by modeling the spatiotemporal relationship between the skeleton joints by a chain graphical structure. They computed the inclination and azimuth angles of joints with respect to the torso basis, as well as the Euclidean distance between joints, to represent pose features. The pose sequence was modeled by a chain graphical structure and aligned by a graph matching technique. The matching energy was used to assess the performance of a given movement sequence aligned to the model sequence.

In recent skeleton-based physical rehabilitation research, temporal alignment between skeleton sequences has been addressed for assessing the quality of physical exercise in several works. Antunes et al. [7] presented a visual and human-interpretable feedback system for guiding the user in the proper performance of certain actions and movements. First, skeleton data were preprocessed for the spatial and temporal alignment between a skeleton instance and the template counterpart. Then, by calculating the Euclidean distance of the 3D coordinates of body parts, the matching error between two skeleton sequences was quantified to assess the quality score. Finally, feedback proposals were automatically computed by minimizing the skeleton matching error.

Baptista et al. [8] employed subsequence dynamic time warping to find the temporal alignment, and the temporal commonality discovery was presented to discover the interval of the subsequence with content that was similar to that of the template action. A qualitative evaluation was performed, and feedback proposals were provided in order to correct the user with respect to the template action.

Paiement et al. [9] studied and compared the performances of different pose representations and HMMs with the dynamics of movement for the online quality assessment of human motion. They developed robust manifold representation and a first-order Markovian assumption to describe the dynamics of a human body pose. Their experimental results showed that the continuous-state HMM was better suited for describing motion dynamics than the classical discrete state HMMs when frame-by-frame deviation was measured for motion quality assessment. In their earlier work of [6], they introduced diffusion maps based on a nonlinear manifold technique to deal with noise and outliers for filtering skeleton data. To assess the quality of movements, they proposed two statistical models to represent the probability density function of the poses and dynamics of the skeleton, and they calculated the log-likelihoods of an observation compared with these models to represent the measures of pose quality and dynamics quality. They tested the proposed method on gait on stairs and demonstrated its generalizability to the movements of unknown subjects and ability to detect a range of abnormality types.

Elkholy et al. [78] proposed to compute handcrafted features from 3D skeletal data provided by depth sensors to detect and assess motion disorders in realistic scenarios. The proposed descriptor consisted of three medical justified features, including asymmetry, velocity magnitude, and Center of Mass (CoM) trajectory deformation, to encode both spatial and temporal characteristics of the motions. Then two probabilistic normalcy models of a Gaussian mixture model (GMM) and Kernel density estimation (KDE) were developed to model normal motion patterns, and the likelihood of a testing sequence was computed and used as a normalcy similarity measure. Furthermore, they trained a multiple linear regression with interactions model for predicting the assessment score to evaluate the

degree of abnormality in performing an action. Experimental results demonstrated that the proposed methods can detect the abnormality in performing different activities for patients with different types of neuromusculoskeletal disorders and assess the quality of the action performance as well.

3.2.2. Handcrafted Feature for Sports Activity Scoring

Some action evaluation studies have addressed automatic sports activity scoring based on the analysis of skeleton data. There are mainly two challenges to be tackled in this field of research. The first one is determining the means of extracting effective and robust features from pose sequences for action quality assessment, and the second one is establishing methods to measure the similarity of pose features to the tolerance of intra-class variations while distinguishing the inter-class confusions. Some research works have regarded action quality assessment as time sequences' modeling and a similarity measurement problem, while others have approached it from the perspective of machine learning.

Gordon [24] first presented the idea of automated assessment of human performance in video. Three key issues concerning the automatic video assessment problem have been addressed: The appropriate performance type, the necessary computer technology, and the way in which automation affects issues surrounding performance scoring. Then, they proposed the use of a motion tracking algorithm in computer vision to extract positions of the human body center from video and compute the difference in positions to assess several components of vaulting performance according to the Federation International de Gymnastique (FIG) Code of Points. Ilg et al. [51] applied an earlier work on constructing an action quality assessment dataset. They used VICON612—a motion capture system equipped with 11 cameras and 41 markers on the human body—to collect 14 karate kata videos performed by seven actors. Then, a hierarchical learning-based method to estimate skill levels from sequences of complex movements in sports was developed. In their learning method, three invariant key features, namely, angular velocity, curvature, and torsion of 3D trajectories, were computed to identify movement primitives first. A sliding window search was performed on a test video to find the subsequence that was the most similar to template features. Then, spatio-temporal morphable models (STMMS) were employed to learn generative models for different styles of movements. Finally, radial basis function (RBF) networks were trained with expert ratings to estimate the skill level of each sequence.

Wnuk and Soatto [79] created the first diving dataset, FINA09, which was gathered from the finals of the diving event at the 13th FINA World Championships. The dataset consists of 68 diving videos performed by 12 different divers with 10 dive types, including 107B, 205B, and 307C. First, they applied background subtraction to obtain foreground regions and Kalman filtering to track the center of mass of the diver. Then, different pose descriptors, action modeling methods, and classification approaches were applied in an attempt to classify the dive types. The experimental results demonstrated that gradient orientation histogram-based pose representation was effective for dive type classification. The SIFT pose features computed after applying the foreground mask within a fixed square window size centered at the diver's tracked location, divided into a 4×4 spatial bin configuration and 8-bin orientation, performed better than those without incorporating background subtraction, as well as HOG-HOF features.

Pirsiavash et al. [1] collected 309 diving and figure skating competition videos to create a new MIT Olympic Scoring dataset. They proposed a general learning-based framework to assess the quality of human-based actions from videos. They extracted both low-level spatiotemporal filtering features that captured edges and velocities, as well as high-level pose features obtained from the discrete cosine transformation (DCT) of joint displacement vectors, and they estimated a regression model that predicted the scores of actions. Compared with the space-time interest points (STIPs) and DFT pose features, the DCT pose features achieved improved rank correlation between the estimated scores and the ground truths for both diving and figure skating.

Venkataraman [3] proposed computing multivariate approximate entropy to model dynamics in individual body joints and cross approximate entropy to model the interaction between body joints.

They used the two-feature representations to quantify the dynamical regularity of human actions for action segmentation and diving action evaluation. In their experiment, the rank correlation was better than that in [1] for diving quality assessment.

3.2.3. Handcrafted Feature for Skill Training

In application to skill training, Sharma [80] applied an earlier attempt to design a fully automated surgical skill assessment framework from video data on the basis of Objective Structured Assessment of Technical Skills (OSATS). They used the STIP detector proposed by Laptev [68] and computed 162-dimensional HOG-HOF descriptors to describe 3D video patches around each detected STIP. These descriptors were clustered by BOW. Then, the sequential motion texture features were computed by a gray-level co-occurrence matrix texture analysis to encode the sequential motion information. The experimental results demonstrated that the combination of local spatial-temporal features with sequential relations helped to improve the performance of automated OSATS assessments.

Relevant studies were furthered in the works of Zia [81,82]. These studies also employed the STIP detector and HOG-HOF descriptor to extract local motion features. The K-means algorithm was used to cluster local features to obtain motion class time series. Then, three types of feature modeling, namely, symbolic features (HMM, BOW, and ABOW), texture features (motion texture and sequential motion texture), and frequency features (DFT and DCT), were applied in an attempt to represent motion features. In the final stage, the nearest-neighbor classifier was trained with the cosine distance metric for sequential feature selection to determine the skill level of surgical actions. They demonstrated that fine-grained analysis of motion dynamics in the frequency domain was more effective than other feature models in representing the skill-relevant information for surgical skill assessment.

Recently, Fard et al. [83] presented a classification framework to automatically assess the expertise level of surgical learners. With the combination of a clinical viewpoint and motion characteristics, they proposed eight global movement features extracted from movement trajectory data, including (task completion time, path length, depth perception, speed, motion smoothness, curvature, turning angle, and tortuosity) extracted from movement trajectory data to quantify the movement pattern of surgical actions. These features were fed into three classification algorithms—k-nearest neighbor (k-NN), logistic regression, and support vector machine—to train a binary classifier that determined two skill levels: Expert and novice.

In summary, the criteria of quality assessment with a diversity of action categories are hard to unify, and some evaluation criteria are not quantifiable. It is difficult to design one mechanical feature engine to extract various kinds of patterns for all-class action evaluation. Furthermore, a handcrafted feature representation has been proved unsuitable for indicating the characteristics of complex activities in long-duration videos. These difficulties lead to large gaps between the estimated results and the ground-truth scores. The evaluation performance of handcrafted feature methods is far from satisfactory.

3.3. Deep Feature Methods for Human Action Evaluation

With the popularity of using convolutional neural networks to tackle the problem of image representation and understanding in computer vision, many studies have taken an interest in developing deep learning methods for the analysis of human activity. Some typical learning networks, such as the 3D convolutional network (C3D) [84,85], long-short term memory (LSTM) [86,87], and the two-stream convolutional network [88,89], have been presented to extract deep features from human activities for action recognition.

The C3D-based feature learning has extended convolutional feature extraction from the spatial dimension to both the spatial and temporal domain. It takes a video sequence as an input with three (x , y , t) dimensions, where x and y represent the position in the spatial domain, and t denotes the coordinates on the temporal axis. Then, three-dimensional convolution is performed in the learning network to extract the semantic spatio-temporal features to represent the dynamic patterns of human

activity. Due to the large-scaled parameters and the high cost of storage capacity, researchers have further proposed decoupling a three-dimensional convolution into two separated branches—spatial and temporal—and named it a two-stream convolution network. One branch of the spatial network takes RGB images as an input and extracts the appearance information of human action, while another branch of the temporal network takes the input of optical images computed from consecutive frames to represent motion information.

With regard to long-duration activity representation in video, there are two main types of approaches: The temporal aggregation methods and temporal relationship learning methods. In temporal aggregated representation, feature extraction is conducted on each frame or segment. Then, a pooling operation, such as max pooling or average pooling, is used to fuse all frame-level or segment-level feature vectors into final video-level features. In temporal relationship learning methods, a recurrent neural network, such as long-short term memory, is employed to model the temporal relationship among the frames or segments and to form the video-level representation.

3.3.1. Deep Features for Physical Rehabilitation

In application to physical therapy or rehabilitation, Parmar [22] performed an earlier comparison of different learning techniques, including support vector machine, single- and double-layered neural networks, and boosted decision trees, to determine whether an action quality is good or bad. They trained a single-layered neural network with 9600 input neurons and 500 neurons in the hidden layer, as well as a two-hidden-layered neural network to classify actions. However, the accuracy of the neural network was inferior to that of the AdaBoost decision tree using the same skeleton feature representation, possibly because of an unreasonable configuration of networks and learning strategy.

Recently, Vakanski et al. [90] created a physical rehabilitation movement dataset called UI-PRMD, which comprises 100 instances of 10 rehabilitation movements captured by a Vicon optical tracker and a Kinect camera. In their work of [91], they further proposed three end-to-end deep learning-based models, namely, CNNs, an RNN with two bidirectional layers of LSTM units, and hierarchical neural networks (HNNs) with five recurrent subnetworks, to encode the relationship between the joint coordinate data and quality assessment scores. Antunes et al. [92] introduced a labeled 3D skeletal dataset called AHA-3D for research on automatic fitness exercise assessment, and it comprises 79 skeleton videos of 21 subjects performing four different movements. They trained a two-stream CNN and an LSTM network separately as the baseline methods for the recognition and detection tasks. The high accuracy rates of 91% in recognition and 88.29% in detection were achieved. In addition, Blanchard et al. [93] presented a dataset consisting of 480 countermovement jumps and drop jumps to facilitate research on assessing anterior cruciate ligament (ACL) injury risk in female athletes from the perspective of computer vision.

3.3.2. Deep Features for Sports Activity Scoring

For sports activity scoring, Parmar also applied a prior attempt to introduce a deep learning method for assessing the action quality of athletes. In the work of [94], they utilized a C3D to extract spatio-temporal features to obtain clip-level feature vectors. Then, the average pooling and LSTM were employed to model the sequential relation between video clips and form the video-level feature representation. Finally, a super vector regression-based scoring estimation function was trained from the input of normalized video-level feature vectors. The experimental results showed that the average pooling performed well in assessing the quality of diving and vaulting videos in a comparison using the LSTM aggregated sequential relation representation. In a further investigation into whether actions have some commonalities among different categories, the AQA-7 dataset [95] was compiled for action quality assessment in sports videos. Using the same C3D-LSTM framework as that used in [94], the authors compared the performance of an all-action model with an action-specific model to test the hypothesis that knowledge transfer is possible in action evaluation. They demonstrated that with the LSTM aggregated video-level feature representation, the all-action model outperformed the

single-action model. It is worth noting that C3D-SVR with the single-action model always achieved the best performance. However, the multi-action model outperformed any single-action model in the evaluation of unseen or novel action classes. Recently, in the work of [96], a multitask learning framework was introduced to tackle three action analysis tasks: Action recognition, commentary generation, and action quality assessment. They proved the effectiveness of C3D-AVG-MTL in capturing the inherent concept of action qualities, and they achieved a 90.44% correlation between the estimated results and the ground-truth scores.

In their research on scoring figure skating activities in video, Xu et al. [97] contributed a figure skating Fis-V dataset that consists of 500 competition videos of 149 professional players and the average length of each video is approximately 170 seconds. They utilized the state-of-the-art deep architecture and proposed a multiscale convolution aggregation model with skip-LSTM for skating activity scoring. The two subnetworks of self-attentive LSTM and multiscale convolutional skip-LSTM were designed in this model: The former was used to learn the different weights of video clips and select important features for total element score regression, and the latter was employed to extract important local and global features while discarding redundant information for a total program component score regression. In [98], Xiang et al. proposed a segmental network called Pseudo-3D (P3D) to tackle the problem of sports activity scoring. They utilized the encoder-decoder temporal convolutional network (ED-TCN) [99] to divide diving videos into four segments: Jumping, dropping, entering the water, and water spray decaying. Then, three types of P3D ResNet networks [100], namely, serial, parallel, and composition of P3D, were used in an attempt to extract feature representation. The last layer of P3D was replaced by a fully connected layer with dropout for regression. They achieved significant improvement on the UNLV-Dive dataset and found that P3D CNN-based feature extraction on full video performed similarly when only water spray was involved. However, the segment-aggregated feature extraction could effectively model the criteria of scoring by weighting the contribution of different stages.

Relevant studies have emerged quickly to diversify this field of research, including approaches such as ScoringNet [101,102] (a 3D CNN network for sports activity scoring) and SwingNet [103] (a hybrid deep neural network and recurrent neural network for assessing golf swing performance), as well as deep learning-based yoga asana recognition [104].

3.3.3. Deep Features for Skill Training

In application to skill training, Wang et al. [105] first explored the development of deep learning methods for assessing surgical skill by analyzing multivariate time-series data of motion kinematics. They proposed a 10-layered convolutional neural network that comprises five types of operations, namely, convolution, pooling, flattening, full connection, and softmax regression, to extract the intrinsic motion dynamics for automatic skill assessment. The network was trained with the loss of the difference between the estimated results and the truth labels by minimizing the multinomial cross-entropy cost. Fawaz et al. [106] designed a CNN method to evaluate surgical skills from kinematic data. In this learning architecture, 76-dimensional kinematic data were captured and inputted into the network. The first layer captured gesture information, and the second layer acquired global information related to surgical skill level. A global average pooling operation was employed after convolution to identify the regions that contributed the most to skill determination, and a fully connected softmax layer predicted the class label (N, I, and E, which denote novice, intermediate, and expert, respectively).

Without the provision of kinematic data or tool motion data from a robotic surgical system, Funke et al. [107] first proposed a video-based surgical skill assessment method using a deep neural network. They fine-tuned a pretrained 3D ConvNet to extract spatio-temporal features from video snippets and used a temporal segment network to resolve ambiguities in single video snippets for the task of surgical skill assessment. Doughty et al. [108] presented a pairwise deep ranking method for skill determination in video and applied the proposed method to a variety of tasks, such as knot tying, needle passing, and suturing in surgery; dough-rolling; drawing; and chopstick-using. Two-stream

CNN was employed to extract the spatio-temporal feature representations for all pairs of videos, and pairwise ranking and similarity were computed in the loss function. In their further study of [109], a new Bristol everyday skill tasks (BEST) dataset was introduced for researching skill determination in long videos. They proposed a rank-aware temporal attention network that comprised I3D feature extraction, attention-filtering representation, and a ranking process, among which multiple attention filters were employed in the attention module to discover important parts in the long video for skill determination. The network was trained with the three types of loss: ranking loss, disparity loss, and rank-aware loss; it achieved state-of-the-art performance on the Epic-Skill and BEST datasets.

A novel RNN-based spatial attention model was proposed in [110] to capture temporal transition patterns of attention combined with high-level task-related features for better skill assessment in video. The learning framework consisted of two RNNs, one for estimating the spatial attention for each frame and the other for skill score estimation. Three processing modules were designed to realize estimation, among which the feature encoding module extracted deep appearance–motion features from RGB and optical flow images. The attention pooling module combined appearance–motion features with task-related features, as well as the temporal relationship between attention. The temporal aggregation module modeled the temporal transition of actions to form the final accumulated representation. The experiments on an infant grasp dataset and BEST dataset [109] demonstrated that considering attention can be helpful for the automatic skill assessment of chopstick-using, dough-rolling, drawing, and surgery.

Although significant progress has been achieved for human action evaluation by deep learning-based feature representation, some key issues deserve to be further studied. On the one hand, the performance on practical application tasks, such as sports activity scoring, physical exercise training, and skill level determination, is below current application requirements. This deficiency is possibly because most published deep learning methods have employed mechanical equal-sized division of long-duration video for the sake of reducing the parameter scale of the learning network. The important temporal information might be lost because of over-segmentation or false segmentation of actions. On the other hand, most of the above-mentioned research works have simply duplicated existing deep feature models designed for action recognition and directly implanted them into the task of action evaluation. Thus, it is difficult to develop a reliable action evaluation method without distinguishing the intrinsic difference between the two problems.

4. Benchmark Datasets and Evaluation Criteria

4.1. Datasets

Large-scale datasets with diverse actions or subjects for validating action evaluation algorithms are still lacking. Most public datasets are oriented toward particular applications, and a limited number of training samples have been collected. As the construction of an action evaluation dataset requires domain experts to provide ground-truth annotations for data, it is difficult to build a universal dataset comprising multiple action categories and massive training data. As an alternative, researchers have created their datasets manually for some specific tasks.

4.1.1. Physical Rehabilitation Datasets

Most physical rehabilitation-relevant datasets remain publicly unavailable because of two major factors. One is the privacy issue concerning patients' or elderly people's private activities, and the other is about the property rights reserved by some medical organizations or funding companies. The five publicly available datasets are introduced dedicated to this field of research: University of Bristol's AI SPHERE-Staircase2014 [111], SPHERE-Walking2015 [112], SPHERE-SitStand2015 [113], the University of Idaho-Physical Rehabilitation Movement Data (UI-PRMD) dataset [114], and the AHA-3D dataset [115]. These publicly available datasets are summarized by the numbers of action

categories, persons, and samples, as well as the action classes and the provision of data modality, as shown in Table 2.

Table 1. The reviewed research works of human action evaluation.

Applications Methods	Physical Therapy	Sports Activity Scoring	Skill Training
Skeleton or kinematic data-based methods	[6–9,22,90–92]	[1–5,48,51,104]	[83,105,106,116–118]
Handcrafted feature learning methods	[6–9,75–78]	[1–5,24,48,49,51,79]	[23,50,80–83,116–118]
Deep feature learning methods	[22,90–93]	[94–98,101–104]	[105–110]

Table 2. Physical rehabilitation dataset.

Dataset Name	#Action Categories	#Persons	#Samples	Data Modality
SPHERE-Staircase2014 dataset [111]	1	12	48	Depth sequences, skeletons
SPHERE-Walking2015 dataset [112]	1	10	40	Depth sequences, skeletons
SPHERE-SitStand 2015 dataset [113]	2	10	109	Depth sequences, skeletons
UI-PRMD dataset [114]	10	10	100	Positions and angles of body joints in the skeletal models
AHA-3D dataset [115]	4	21	79	3D skeletal sequences, RGB images

Paiement et al. [6,9] have devoted themselves to the motion quality assessment of physical therapy and rehabilitation. They presented the SPHERE-Staircase2014, SPHERE-Walking2015, and SPHERE-SitStand2015 datasets successively for evaluating action quality assessment methods that can be used to support the physical activity of stroke or sports injury patients at home. The SPHERE-Staircase2014 dataset [6] includes 48 sequences of walking upstairs performed by 12 individuals captured by an Asus Xmotion RGB-D camera placed at the top of the stairs in a frontal and downward-looking position. The SPHERE-Walking2015 dataset [9] includes 40 sequences of 10 individuals walking on a flat surface. This dataset was captured by an Asus Xmotion RGB-D camera placed in front of the subject. It contains normal gaits and two types of abnormal gaits to simulate stroke and Parkinson disease patients walking under the guidance of a physiotherapist. The SPHERE-SitStand2015 dataset [9] includes 109 sequences of 10 individuals performing two actions—sitting down and standing up—in a home environment. Instead of the Asus Xmotion RGB-D camera, a Kinect V2 camera was used to track the skeleton for movements. Two types of abnormal movements and normal motions are included.

Vakanski et al. [90] presented the UI-PRMD dataset of human body movements for physical rehabilitation exercise. They used a Vicon optical tracker and a Microsoft Kinect sensor to capture the positions and angles of full-body joints. For data collection, 10 subjects were recruited to perform 10 different movements repetitively, including the deep squat, hurdle step, inline lunge, side lunge, sit to stand, standing active straight leg raise, standing shoulder abduction, standing shoulder extension, standing shoulder internal–external rotation, and standing shoulder scaption.

Recently, the AHA-3D dataset [92] was collected for assessing the fitness level of seniors during exercise. This dataset provides 3D skeletal data acquired from Kinect v2 and color images from RGB cameras. A sample of 21 subjects consisting of 11 young individuals and 10 elderly individuals (5 males and 16 females) performed four different motions: 30-second chair stand, 8-foot up and go, 2-minute step test, and unipedal stance. The result was the capture of 79 skeleton videos.

4.1.2. Sports Activity Scoring Datasets

In the research on automatic sports activity scoring, several sports activity datasets have been collected from international competition events on diving, figure skating, vaulting, and others. Several publicly available benchmark datasets are introduced, including the FINA09 diving dataset [79], MIT Olympic scoring dataset [1], UNLV AQA-7 dataset [95], MTL-AQA dataset [96], and Fis-V dataset [97], etc. The details of these benchmarks are presented in Table 3 from the perspective of action categories, the number of samples, data source, viewpoint, and variation in the background.

Table 3. Sports activity scoring dataset.

Dataset Name	#Action Categories	#Samples	View of Samples	Background of Samples
FINA09 diving dataset [79]	1	68	Front, side	Same
MIT Olympic Scoring dataset [1]	2	309	Variations	Same
UNLV AQA-7 dataset [95]	7	1189	Severe changes	Different
MTL-AQA dataset [96]	1	1412	Severe changes	Different
Fis-V dataset [97]	1	500	Severe changes	Different
GolfDB dataset [119]	1	1400	Multiple views	Different
YogaVidCollected dataset [120]	6	88	Small changes	Same

The FINA09 diving dataset [79] was collected from the 13th FINA World Championships of diving, and it contains 68 annotated diving videos in total, with 12 athletes each performing about six unique dives. Each dive was recorded from front and side viewpoints. The ground-truth score of each dive was obtained by the product of the execution score multiplied by the difficulty score.

The MIT Olympic Scoring dataset [1] is made of 159 diving videos and 150 figure skating videos captured from the Olympic competition and World Championship events on YouTube. The frame rate of diving is 60 frames per second (fps), about 150 frames per video, and the AQA score ranges from 20 (worst) to 100 (best). The frame rate of the figure skating video is 24 fps, about 4200 frames per video, and the AQA score ranges from 0 (worst) to 100 (best). The front and side viewpoints provided for diving, while severe changes of view exist in figure skating.

The UNLV AQA-7 dataset [95] comprises 1189 videos of seven action categories: 370 singles diving-10m platform, 88 synchronous diving-3m springboard, 91 synchronous diving-10m platform, 176 gymnastic vaulting, 175 skiing, 206 snowboarding, and 83 trampoline videos were captured from Summer and Winter Olympics events on YouTube. There is a difference in the average length of videos among the seven action categories, ranging from 87 (vaulting) to 634 (trampoline). Severe changes in view exist in vaulting, skiing, and snowboarding, while small and negligible variations in viewpoint exist in trampoline and diving videos. The ground-truth score of each action is obtained by the product of the execution score multiplied by the difficulty score.

The MTL-AQA dataset [96], the largest diving dataset to date, was collected from 16 international competition events and contains 1412 samples of single or synchronized diving performances from 10 m platform and 3 m springboard diving. The variations in this challenging dataset include viewpoint, male and female divers, and cluttered background. Each sample is annotated with three labels: the action quality assessment score, action category, and commentary.

Xu et al. [97] proposed the Fis-V dataset, which comprises 500 figure skating videos collected from high-standard international skating competitions, including the NHK Trophy (NHK), Trophée Eric Bompard (TEB), the Cup of China (COC), and the Four Continents Figure Skating Championships (4CC). The frame rate is 25 fps, about 4300 frames per video, performed by 149 athletes from 20 different countries. Each video is annotated by two scores: Total element score (TES) and total program component score (PCS).

To facilitate golf swing analysis, McNally et al. [103] introduced GolfDB, a benchmark video database for the novel task of golf swing sequencing. It was collected from 580 YouTube videos and consists of professional golfers from the PGA, LPGA, and Champions Tours, totaling 248 individuals

with diverse golf swings. A total of 1400 high-quality golf swing videos and over 390k frames captured from multiple camera angles and multiple backgrounds were extracted.

Since yoga exercises provide physical, mental, and spiritual benefits to people, yoga learning and self-instruction systems are becoming popular across the globe. Some researchers have investigated computer-assisted self-training systems for improving the performance of yoga exercises. Recently, Yadav et al. developed a deep learning algorithm to recognize various yoga asanas. To this end, a yoga dataset called YogaVidCollected [120] was created using an HD 1080p RGB webcam that recorded 15 individuals performing six yoga asanas (i.e., Bhujangasana, Padmasana, Shavasana, Tadasana, Trikonasana, and Vrikshasana). In total, 88 videos with approximately 111,750 frames were collected in an indoor environment, and most were performed at a 4–5-meter distance from the front of the camera.

4.1.3. Skill Training-Related Datasets

In research on skill training, some datasets have been created for evaluating the action performance of expertise learners for particular tasks, such as surgical skill training and monitoring and assisting the daily lives of elderly people. Several publicly available datasets have been introduced in this application area, including the JIGSAWS dataset [121–123], EPIC-Skills 2018 dataset [124], BEST 2019 dataset [125], breakfast action dataset [126], and the ADL dataset [127], as illustrated in Table 4.

Table 4. Skill training-related dataset.

Dataset Name	#Action Categories	#Samples	Data Modality	View of Samples	Background of Samples
JIGSAWS dataset [122]	3	103	Kinematics data, video data	two left and right cameras	Same
EPIC-Skills 2018 dataset [124]	6	216	Video data	single view	Different
BEST 2019 dataset [125]	5	500	Video data	Severe changes	Different
Breakfast Actions database [126]	10	1989	Video data	3–5 cameras	Different
ADL dataset [127]	18	440	Video data	170-degree first-person view angle	Different

The JIGSAWS dataset [122,123] was captured by a da Vinci Surgical System and consists of kinematics and video data from eight participants for three robotic surgical tasks: Suturing (SU), knot tying (KT), and needle passing (NP). All subjects repeated each surgical task five times. As some trials (1 for SU, 4 for KT, and 12 for NP) were unusable as a result of corrupted data recordings, the final dataset consists of 39 samples of SU, 36 samples of KT, and 28 samples of NP. This dataset was collected for the Language of Surgery project at the John Hopkins University site. The project is dedicated to developing models for analyzing surgical motion data and applying these models to teaching and assessing action performance of surgical trainees.

The EPIC-Skills 2018 dataset [108,124] comprises 216 samples of six tasks from two published and two newly recorded datasets: 103 samples of knot tying, needle passing, and suturing from JIGSAWS are used for surgical skill assessment, and 33 dough-rolling samples performed by 33 persons from CMU-MMAC dataset are used for assessing the pizza making activity. Another two tasks, drawing and chopstick-using, were captured by a high-resolution camera with a single position and background.

Since the limitations of the EPIC-Skills dataset of a single view and scarcely cluttered background result in scenes that are far from realistic of applications, a novel skill assessment dataset—Bristol everyday skill tasks (BEST) 2019 [109,125]—was collected. The BEST dataset comprises 500 videos of five skill tasks: Scrambling eggs, braiding hair, tying a tie, making an origami crane, and applying eyeliner. The videos were retrieved and compiled from YouTube. Therefore, the recording view and background may differ. Each video is annotated with a class label of B, I, or E, which indicate beginner,

intermediate, and expert, respectively. Then, pairwise annotation is performed on 40% of the possible pairs for skill ranking.

Another practical application of a skill training system is to provide living assistance for elderly people, protecting them from dangers or accidents in their daily life. Several datasets involving daily activities, such as breakfast preparation and cooking, have been created, for example, the Breakfast Actions dataset [126], ADL dataset [127], MPII cooking dataset [128], Charades [129], and EPIC-KITCHEN [130]. The breakfast actions, MPII cooking, and epic-kitchen datasets are recorded cooking activities. The ADL and Charades datasets contain hundreds of action classes and thousands of video sequences involving human daily lives in order to facilitate a computer vision system-based analysis of human daily activities.

4.2. Performance Evaluation Criteria

In the published research, non-uniform evaluation criteria have been adopted for validating action evaluation algorithms since different problem definitions have been formulated in different application fields. In physical rehabilitation, two-class classification accuracy, namely, normal and abnormal movements, is commonly used as the evaluation criteria. A regression problem is generally formulated in sports activity scoring. Thus, the precision of the regression that measures the similarity between the prediction scores and ground-truth scores is determined to verify action quality assessment algorithms. Nevertheless, the ranking accuracy, classification, and regression are all used for validating the effectiveness of skill assessment methods.

As physical rehabilitation entails private activities and property rights restrictions, a few datasets in this field are publicly available. The published research works have employed diverse evaluation methodologies and verified the proposed algorithms on their own datasets. Therefore, it is difficult to compare them using a unified criterion. Paiement et al. [6,9] formulated an online quality assessment for abnormal events detection. They developed a statistical normal movement model and measured how much a movement deviates from normal. They achieved 84% of abnormal events detection of gait on stairs on the SPHERE-Staircase2014 dataset, and the best AUCs were 1.00, 0.99, and 1.00 for walking, sitting, and standing on the SPHERE-Walking2015 and SPHERE-SitStand2015 datasets with the selected model parameters. Elkholy et al. [78] developed two probabilistic normalcy models of a Gaussian mixture model (GMM) and Kernel density estimation (KDE) to assess physical movements, and the best AUCs achieved were 1.00 for walking, sitting, and standing on the SPHERE datasets using their GMM model. With their KDE model, the best AUCs were 1.00 for walking, sitting, and standing and 0.98 for gait on stairs. They also employed an equal error rate, detection rate at a 1% false acceptance rate, and detection rate at a 5% false acceptance rate to evaluate the performance of the proposed method. Liao and Vakanski [90,91] introduced the UI-PRMD dataset and developed a deep learning framework for assessing physical rehabilitation exercises. They used and compared the regression performance of three deep learning networks—CNN, RNN, and HNN—as well as four metrics: Euclidean distance, Mahalanobis distance, DTW distance, and GMM log-likelihood.

In sports activity scoring, the performance of an action quality assessment method is commonly evaluated by the similarity between the predicted results and the ground-truth scores. The Spearman rank correlation coefficient or Pearson's correlation coefficient is commonly employed to measure the similarity: The higher the value, the better the performance. The performance of some state-of-the-art methods on the MIT Olympic Scoring dataset and UNLV AQA-7 dataset are presented in Table 5. A comparison of performances on the MIT dataset makes it apparent that deep learning methods [94,95,97,98,101,102] significantly improve the performance of sports scoring estimation compared with handcrafted approaches [1,3]. The best performance on MIT diving improved to 0.86, and the figure skating was 0.59. With a larger dataset, UNLV AQA-7, which has more than one thousand videos, the best rank correlations are 0.84 and 0.7 for the diving and vaulting videos, respectively.

Table 5. Action evaluation performance of methods on sports scoring dataset. The superscript D indicates a deep learning method.

Methods	Year	MIT Olympic Scoring Diving	MIT Olympic Scoring Skating	UNLV AQA-7 Diving	UNLV AQA-7 Vault
[1]	2014	0.41	0.35		
[3]	2015	0.45			
[94] ^D	2017	0.74	0.53	0.79	0.68
[95] ^D	2018			0.61	0.67
[101] ^D	2018		0.57	0.80	0.70
[102] ^D	2018	0.78		0.84	0.70
[98] ^D	2018	0.86			
[97] ^D	2019		0.59		

The skill assessment approaches commonly employ three types of evaluating criteria: Classification accuracy, score estimation accuracy, and rank accuracy. For example, three expertise levels, namely, expert, intermediate, and novice, are defined in the JIGSAWS dataset, and the classification rate of accurate predictions the class label (E, I, N) is regarded as the classification accuracy. Most published research works [83,105,116–118] report their classification accuracy as detailed in Table 6. A small selection of works, except for [118], have reported the correlation coefficient between the prediction scores and ground-truths as score estimation accuracy. Some researchers have suggested that the ranking of videos is more suitable than estimating an objective score [108,110]. Therefore, rank accuracy, which is defined as the percentage of correctly ordered videos in a ranking, has been employed to evaluate skill determination algorithms.

Table 6. Action evaluation performance of methods on the JIGSAWS dataset. The superscript D indicates a deep learning method.

Method	Year	Evaluation Criteria	Action Categories		
			Suturing	Knot Tying	Needle Passing
[116]	2012	Classification accuracy	97.4%	96.2%	94.4%
[117]	2017	Classification accuracy	89.7%	96.3%	61.1%
[118]	2018	Classification accuracy	100%	99.9%	100%
		Score prediction (Correlation Coefficient)	0.75	0.63	0.46
[83]	2018	Classification accuracy	89.9%	95.8%	82.3%
[105] ^D	2018	Classification accuracy	93.4%	89.8%	84.9%
[107] ^D	2019	Classification accuracy	100%	-	96.4%
[108] ^D	2018	Rank accuracy		70.2%	
[110] ^D	2019	Rank accuracy		73.1%	

5. Conclusions and Discussion

In this paper, a comprehensive review was carried out of recent computer vision-based approaches for human action evaluation research. The relevant literature was summarized according to several key issues of action evaluation research, including motion representation based on skeleton data detection and preprocessing, handcrafted feature methods, and deep feature methods in different real-world applications. The benchmark datasets collected from the field of physical rehabilitation, sports activity scoring, and skill training were described. Useful information was provided to facilitate the researchers' selection of suitable datasets for their studies. Furthermore, the effects of published action evaluation research were discussed by comparing their feature learning and assessing methods, as well as their performance on several benchmark datasets. The authors conclude the study on human action evaluation research with the following observations:

- (1) Most existing research works have directly employed traditional machine learning or state-of-the-art deep learning methods in the field of action recognition to tackle the problem of

action evaluation. Thus, it is difficult to develop reliable action quality assessment algorithms without distinguishing the intrinsic difference between these two problems.

- (2) In the application of sports activity scoring, deep feature representation methods have been proved superior than handcrafted feature methods in their performance on benchmark datasets. Specifically, most of deep-learning methods significantly improve the scoring estimation results on MIT Olympic Scoring dataset than handcrafted approaches. The best correlation coefficient of dive scoring has been improved to 0.86, and the figure skating is 0.59 as introduced in Table 5. The reason is probably because that it is difficult to design one mechanical feature engine to extract various kinds of patterns for multiple-class action evaluation. Therefore, handcrafted feature representation is unsuitable for indicating characteristics of complex activities in long-duration video. On the other hand, most of deep learning methods have employed mechanical equal-sized division on long-duration video for the sake of reducing the parameter scale of learning networks. The important temporal information might be lost as a result of over-segmentation or false segmentation.
- (3) In skill training, both the handcrafted and deep learning feature methods achieved high classification accuracy on the JIGSAWS dataset as presented in Table 6. However, a three-category of (E, I, N) classification evaluation is rather simple to evaluate the different methods. More appropriate evaluation criteria, such as rank accuracy and score prediction, deserves further investigation. In physical rehabilitation, a few datasets have been publicly available due to private activities and property rights. The reviewed studies employed diverse evaluation criteria, such as an abnormal events detection rate, equal error rate, and a detection rate with a false acceptance rate, on their own datasets. It is difficult to compare them in a unified criterion. The pro and cons of different methods remains to be further observation.
- (4) The deep feature representation methods have significantly improved the performance on several benchmark datasets. However, their accuracy and efficiency are far from satisfactory and below the current application requirements.
- (5) There is still a lack of large-scale annotated datasets with a diversity of action categories and application fields. This is mainly because of the great labor cost of the domain experts' professional annotation. There is also a lack of unified evaluation criteria to validate the effectiveness of the proposed methods.

Although the difficulties and challenges remain in action evaluation research, such as intra-class variation in diversity and complexity of human movement, camera motion, view changes, occlusion, and background clutter, the performance of deep learning-based feature representation has improved. There are still some key issues that must be addressed for significant progress.

- (1) It is reasonably assumed that the quality of human actions is directly determined by the dynamic change in human body movement rather than environmental factors. Thus, accurate skeleton data detection and deep feature representation methods based on skeleton data are the key issues in the development of reliable quality assessment algorithms for human action evaluation.
- (2) The segmentation of long-duration video on the basis of primitive action semantics and the representation of temporal relationships between action segments are important topics of future deep architecture research for human action evaluation. Most previously published deep learning methods have employed an equal-sized division of video to reduce the parameter scale of learning networks. The important temporal information might be lost as a result of over-segmentation or false segmentation.
- (3) The semantic granularity of the evaluation models needs to be further studied. Most existing studies have adopted a unified regression function to assess all action categories. Thus, the evaluation accuracy has significantly decreased under circumstances of unequally distributed training samples. Furthermore, an all-action regression model is not capable of assessing the quality of unseen actions. Whether a specific-action model or an all-action model is more suitable

for evaluating the quality of actions and whether knowledge transfer can be adapted to train a unified evaluating model across action categories are promising directions that deserve study.

Author Contributions: J.-X.D. reviewed skeleton-based motion detection and preprocessing methods of human action evaluation. Q.L. and H.-B.Z. reviewed the deep learning methods of human action evaluation. S.Y. and D.-S.C. reviewed the handcrafted feature methods of human action evaluation. All authors read and approved the final manuscript.

Funding: This work is supported by the National Nature Science Foundation of China (Grant no. 61673186, 61871196), the Natural Science Foundation of Fujian Province, China (Grant no. 2019J01082, 2017J01110), and the Scientific Research Funds of Huaqiao University, China (16BS812).

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable and insightful comments on an earlier version of this manuscript.

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

References

1. Pirsiavash, H.; Vondrick, C.; Torralba, A. Assessing the Quality of Actions. In *Proceedings of the European Conference on Computer Vision 2014*; Springer: Cham, Switzerland, 2014; pp. 556–571.
2. Patrona, F.; Chatzitofis, A.; Zarpalas, D.; Daras, P. Motion analysis: Action detection, recognition and evaluation based on motion capture data. *Pattern Recognit.* **2018**, *76*, 612–622. [[CrossRef](#)]
3. Venkataraman, V.; Vlachos, I.; Turaga, P. Dynamical Regularity for Action Analysis. In *26th British Machine Vision Conference*; British Machine Vision Association: Swansea, Wales, 2015; pp. 67.1–67.12.
4. Weeratunga, K.; Dharmaratne, A.; How, K.B. Application of Computer Vision and Vector Space Model for Tactical Movement Classification in Badminton. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, 21–26 July 2017; pp. 132–138.
5. Morel, M.; Kulpa, R.; Sorel, A. Automatic and Generic Evaluation of Spatial and Temporal Errors in Sport Motions. In *Proceedings of the International Conference on Computer Vision Theory and Applications*, Rome, Italy, 27–29 February 2016; pp. 542–551.
6. Paiement, A.; Tao, L.; Hannuna, S. Online quality assessment of human movement from skeleton data. In *Proceedings of the British Machine Vision Conference (BMVC 2014)*, Nottingham, UK, 1–5 September 2014; pp. 153–166.
7. Antunes, M.; Baptista, R.; Demisse, G.; Aouada, D.; Ottersten, B. Visual and Human-Interpretable Feedback for Assisting Physical Activity. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, 11–14 October 2016; pp. 115–129.
8. Baptista, R.; Antunes, M.; Aouada, D. Video-Based Feedback for Assisting Physical Activity. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, Rome, Italy, 27 February–1 March 2017.
9. Tao, L.; Paiement, A.; Damen, D. A comparative study of pose representation and dynamics modelling for online motion quality assessment. *Comput. Vis. Image Underst.* **2016**, *148*, 136–152. [[CrossRef](#)]
10. Meng, M.; Drira, H.; Boonaert, J. Distances evolution analysis for online and off-line human object interaction recognition. *Image Vis. Comput.* **2018**, *70*, 32–45. [[CrossRef](#)]
11. Zhang, W.; Liu, Z.; Zhou, L.; Leung, H.; Chan, A.B. Martial arts, dancing and sports dataset: A challenging stereo and multi-view dataset for 3d human pose estimation. *Image Vis. Comput.* **2017**, *61*, 22–39. [[CrossRef](#)]
12. Laraba, S.; Tilmanne, J. Dance performance evaluation using hidden markov models. *Comput. Animat. Virtual Worlds* **2016**, *27*, 321–329. [[CrossRef](#)]
13. Barnachon, M.; Boufama, B.; Guillou, E. A real-time system for motion retrieval and interpretation. *Pattern Recognit. Lett.* **2013**, *34*, 1789–1798. [[CrossRef](#)]
14. Hu, M.C.; Chen, C.W.; Cheng, W.H.; Chang, C.H.; Lai, J.H.; Wu, J.L. Real-time human movement retrieval and assessment with kinect sensor. *IEEE Trans. Cybern.* **2014**, *45*, 742–753. [[CrossRef](#)]
15. Liu, X.; He, G.F.; Peng, S.J.; Cheung, Y.M.; Tang, Y.Y. Efficient human motion retrieval via temporal adjacent bag of words and discriminative neighborhood preserving dictionary learning. *IEEE Trans. Hum. Mach. Syst.* **2017**, *47*, 763–776. [[CrossRef](#)]

16. Girdhar, R.; Ramanan, D.; Gupta, A.; Sivic, J.; Russell, B. Actionvlad: Learning spatio-temporal aggregation for action classification. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 971–980.
17. Wang, H.; Oneata, D.; Verbeek, J.; Schmid, C. A robust and efficient video representation for action recognition. *Int. J. Comput. Vis.* **2016**, *119*, 219–238. [[CrossRef](#)]
18. Duarte, K.; Rawat, Y.S.; Shah, M. Videocapsulenet: A simplified network for action detection. Proceedings of Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 7610–7619.
19. Zolfaghari, M.; Singh, K.; Brox, T. Eco: Efficient convolutional network for online video understanding. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 695–712.
20. Vondrick, C.; Pirsaviash, H.; Torralba, A. Anticipating visual representations from unlabeled video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vegas Valley, NV, USA, 26 June–1 July 2016; pp. 98–106.
21. Becattini, F.; Uricchio, T.; Seidenari, L.; Bimbo, A.D.; Ballan, L. Am I done? Predicting action progress in videos. *arXiv* **2017**, arXiv:1705.01781.
22. Parmar, P.; Morris, B. Measuring the quality of exercises. In Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), Orlando, FL, USA, 16–20 August 2016; pp. 2241–2244.
23. Zia, A.; Sharma, Y.; Bettadapura, V.; Sarin, E.L.; Clements, M.A.; Essa, I. Automated assessment of surgical skills using frequency analysis. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 430–438.
24. Gordon, A.S. Automated video assessment of human performance. In Proceedings of the 7th World Conference on Artificial Intelligence in Education (AI-ED 1995), Washington, DC, USA, 16–19 August 1995; pp. 16–19.
25. Atiqur Rahman Ahad, M.; Das Antar, A.; Shahid, O. Vision-based Action Understanding for Assistive Healthcare: A Short Review. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2019, Long Beach, CA, USA, 15–21 June 2019; pp. 1–11.
26. Aggarwal, J.K.; Ryoo, M.S. Human activity analysis: A review. *ACM Comput. Surv. (CSUR)* **2011**, *43*, 16. [[CrossRef](#)]
27. Yu, K.; Yun, F. Human Action Recognition and Prediction: A Survey. *arXiv* **2018**, arXiv:1806.11230.
28. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **2017**, *60*, 4–21. [[CrossRef](#)]
29. Ziaefard, M.; Bergevin, R. Semantic human activity recognition: A literature review. *Pattern Recognit.* **2015**, *48*, 2329–2345. [[CrossRef](#)]
30. 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](#)] [[PubMed](#)]
31. POPPE, R. A survey on vision-based human action recognition. *Image Vis. Comput.* **2010**, *28*, 976–990. [[CrossRef](#)]
32. Zhu, F.; Shao, L.; Xie, J.; Fang, Y. From handcrafted to learned representations for human action recognition: A survey. *Image Vis. Comput.* **2016**, *55*, 42–52. [[CrossRef](#)]
33. Guo, G.; Lai, A. A survey on still image based human action recognition. *Pattern Recognit.* **2014**, *47*, 3343–3361. [[CrossRef](#)]
34. Alexander, K.; Marszalek, M.; Schmid, C. A Spatio-Temporal Descriptor Based on 3D-Gradients. In Proceedings of the British Machine Vision Conference 2008, Leeds, UK, 1–4 September 2008.
35. Liu, J.; Kuipers, B.; Sararese, S. Recognizing human actions by attributes. In Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3337–3344.
36. Neibles, J.C.; Chen, C.W.; Li, F.F. Modeling temporal structure of decomposable motion segments for activity classification. In Proceedings of the European Conference on Computer Vision 2010, Heraklion, Greece, 5–11 September 2010; pp. 392–405.
37. Shu, Z.; Yun, K.; Samaras, D. Action Detection with Improved Dense Trajectories and Sliding Window. In *Proceedings of ECCV 2014*; Springer: Cham, Switzerland; pp. 541–551.

38. Oneata, D.; Verbeek, J.J.; Schmid, C. Efficient Action Localization with Approximately Normalized Fisher Vectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014, Columbus, OH, USA, 23–28 June 2014; pp. 2545–2552.
39. Shou, Z.; Wang, D.; Chang, S.F. Temporal action localization in untrimmed videos via multi-stage cnns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Vegas Valley, NV, USA, 26 June–1 July 2016; pp. 1049–1058.
40. Yu, G.; Yuan, J. Fast action proposals for human action detection and search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015, Boston, MA, USA, 7–12 June 2015; pp. 1302–1311.
41. Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; Lin, D. Temporal action detection with structured segment networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2017, Venice, Italy, 22–29 October 2017; pp. 2914–2923.
42. Kong, Y.; Fu, Y. Max-margin heterogeneous information machine for RGB-D action recognition. *Int. J. Comput. Vis.* **2017**, *123*, 350–371. [[CrossRef](#)]
43. Hu, J.F.; Zheng, W.S.; Ma, L.; Wang, G.; Lai, J.H.; Zhang, J. Early action prediction by soft regression. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *1*. [[CrossRef](#)]
44. Martinez, J.; Black, M.J.; Romero, J. On human motion prediction using recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 2891–2900.
45. Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Savarese, S. Social lstm: Human trajectory prediction in crowded spaces. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Vegas Valley, NV, USA, 26 June–1 July 2016; pp. 961–971.
46. Xu, H.; Gao, Y.; Yu, F.; Darrell, T. End-to-end learning of driving models from large-scale video datasets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 2174–2182.
47. Kuefler, A.; Morton, J.; Wheeler, T.; Kochenderfer, M. Imitating driver behavior with generative adversarial networks. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV 2017), Los Angeles, CA, USA, 11–14 June 2017; pp. 204–211.
48. Alexiadis, D.S.; Kelly, P.; Daras, P.; OConnor, N.E.; Boubekeur, T.; Moussa, M.B. Evaluating a dancer’s performance using kinect-based skeleton tracking. In Proceedings of the 19th ACM international conference on Multimedia ACM 2011, Scottsdale, AZ, USA, 28 November–1 December 2011; pp. 659–662.
49. Jug, M.; Pers, J.; Dezman, B.; Kovacic, S. Trajectory based assessment of coordinated human activity. In *International Conference on Computer Vision Systems 2003*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 534–543.
50. Reiley, C.E.; Lin, H.C.; Yuh, D.D.; Hager, G.D. Review of methods for objective surgical skill evaluation. *Surg. Endosc.* **2011**, *25*, 356–366. [[CrossRef](#)] [[PubMed](#)]
51. Ilg, W.; Mezger, J.; Giese, M. Estimation of skill levels in sports based on hierarchical spatio-temporal correspondences. In *Joint Pattern Recognition Symposium 2003, Magdeburg, Germany, 10–12 September 2003*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 523–531.
52. 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](#)] [[PubMed](#)]
53. Tran, D.; Bourdev, L.D.; Fergus, R.; Torresani, L.; Paluri, M. C3D: Generic features for video analysis. *arXiv* **2014**, arXiv:1412.0767.
54. Shi, X.J.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Proceedings of the Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 802–810.
55. Ng, J.Y.H.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
56. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.

57. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D. Cascade object detection with deformable part models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2010), San Francisco, CA, USA, 13–18 June 2010; pp. 2241–2248.
58. Yang, Y.; Ramanan, D. Articulated Pose Estimation with Flexible Mixtures of Parts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1385–1392.
59. 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.
60. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the 30th IEEE Conference Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1302–1310.
61. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. Rmpe: Regional multi-person pose estimation. In Proceedings of the IEEE International Conference on Computer Vision 2017, Venice, Italy, 22–29 October 2017; pp. 2334–2343.
62. Guler, 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 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7297–7306.
63. Han, F.; Reily, B.; Hoff, W.; Zhang, H. Space-time representation of people based on 3d skeletal data: A review. *Comput. Vis. Image Underst.* **2017**, *158*, 85–105. [[CrossRef](#)]
64. Pazhoumand-Dar, H.; Lam, C.P.; Masek, M. Joint movement similarities for robust 3d action recognition using skeletal data. *J. Vis. Commun. Image Represent.* **2015**, *30*, 10–21. [[CrossRef](#)]
65. Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; Bajcsy, R. Sequence of the Most Informative Joints (SMIJ): A new representation for human skeletal action recognition. *J. Vis. Commun. Image Represent.* **2014**, *25*, 24–38. [[CrossRef](#)]
66. 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](#)]
67. Amor, B.B.; Su, J.; Srivastava, A. Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1–13. [[CrossRef](#)]
68. Dollar, P.; Rabaud, V.; Cottrell, G. Behaviour recognition via sparse spatio-temporal features. In Proceedings of the 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, 15–16 October 2005; pp. 65–72.
69. Laptev, I.; Lindeberg, T. On Space-time interest points. In Proceedings of the International Conference on Computer Vision 2003, Nice, France, 14–17 October 2003; pp. 432–439.
70. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning realistic human actions from movies. In Proceedings of the Conference on Computer Vision and Pattern Recognition 2008, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
71. Scovanner, P.; Ali, S.; Shah, M. A 3-dimensional SIFT descriptor and its application to action recognition. In Proceedings of the International Conference on Multimedia 2007, Augsburg, Germany, 24–29 September 2007; pp. 357–360.
72. Wang, H.; Schmid, C. Action Recognition with Improved Trajectories. In Proceedings of the IEEE International Conference on Computer Vision 2013, Sydney, Australia, 1–8 December 2013; pp. 3551–3558.
73. Csurka, G.; Dance, C.; Fan, L. Visual Categorization with Bags of Keypoints. In Proceedings of the Workshop on Statistical Learning in Computer Vision (ECCV), Prague, Czech Republic, 11–14 May 2004; pp. 1–22.
74. Vicente, I.S.; Kyrki, V.; Kragic, D.; Larsson, M. Action recognition and understanding through motor primitives. *Adv. Robot.* **2007**, *21*, 1687–1707. [[CrossRef](#)]
75. Chen, Y.; Duff, M.; Lehrer, N.; Sundaram, H.; He, J.; Wolf, S.L.; Rikakis, T. A computational framework for quantitative evaluation of movement during rehabilitation. *AIP Conf. Proc.* **2011**, *1371*, 317–326.
76. Venkataraman, V.; Turaga, P.; Lehrer, N.; Baran, M.; Rikakis, T.; Wolf, S. Attractor-shape for dynamical analysis of human movement: Applications in stroke rehabilitation and action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2013, Portland, OR, USA, 23–28 June 2013; pp. 514–520.

77. Çeliktutan, O.; Akgul, C.B.; Wolf, C.; Sankur, B. Graph-based analysis of physical exercise actions. In Proceedings of the 1st ACM international workshop on Multimedia indexing and information retrieval for healthcare 2013, Barcelona, Spain, 22 October 2013; pp. 23–32.
78. Elkholy, A.; Hussein, M.; Gomaa, W.; Damen, D.; Saba, E. Efficient and Robust Skeleton-Based Quality Assessment and Abnormality Detection in Human Action Performance. *IEEE J. Biomed. Health Inform.* **2019**. [[CrossRef](#)] [[PubMed](#)]
79. Wnuk, K.; Soatto, S. Analyzing diving: A dataset for judging action quality. In *Asian Conference on Computer Vision, Queenstown, New Zealand, 8–12 November 2010*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 266–276.
80. Sharma, Y.; Bettadapura, V.; Plotz, T.; Hammerla, N.; Mellor, S.; McNaney, R.; Olivier, P.; Deshmukh, S.; McCaskie, A.; Essa, I. Video based assessment of OSATs using sequential motion textures. In Proceedings of the International Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI)-Workshop, Boston, MA, USA, 14–18 September 2014.
81. Zia, A.; Sharma, Y.; Bettadapura, V.; Sarin, E.L.; Ploetz, T.; Clements, M.A.; Essa, I. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *Int. J. Comput. Assist. Radiol. Surg.* **2016**, *11*, 1623–1636. [[CrossRef](#)] [[PubMed](#)]
82. Zia, A. Automated Benchmarking of Surgical Skills Using Machine Learning. Ph.D. Thesis, Georgia Institute of Technology, Atlanta, GA, USA, 2018.
83. Fard, M.J.; Ameri, S.; Darin Ellis, R.; Chinnam, R.B.; Pandya, A.K.; Klein, M.D. Automated robot-assisted surgical skill evaluation: Predictive analytics approach. *Int. J. Med. Robot. Comput. Assist. Surg.* **2018**, *14*, e1850. [[CrossRef](#)] [[PubMed](#)]
84. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Washington, DC, USA, 7–13 December 2015; pp. 4489–4497.
85. Sun, L.; Jia, K.; Yeung, D.Y.; Shi, B.E. Human action recognition using factorized spatio-temporal convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision 2015, Santiago, Chile, 7–13 December 2015; pp. 4597–4605.
86. 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 2015, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
87. Li, Z.; Gavriluyk, K.; Gavves, E.; Jain, M.; Snoek, C.G. Videolstm convolves, attends and flows for action recognition. *Comput. Vis. Image Underst.* **2018**, *166*, 41–50. [[CrossRef](#)]
88. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the Neural Information Processing Systems, Motreal, QC, Canada, 8–13 December 2014; pp. 568–576.
89. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, Vegas Valley, NV, USA, 26 June–1 July 2016; pp. 1933–1941.
90. Vakanski, A.; Jun, H.P.; Paul, D.; Baker, R. A data set of human body movements for physical rehabilitation exercises. *Data* **2018**, *3*, 2. [[CrossRef](#)]
91. Liao, Y.; Vakanski, A.; Xian, M. A Deep Learning Framework for Assessing Physical Rehabilitation Exercises. *arXiv* **2019**, arXiv:1901.10435.
92. Antunes, J.; Bernardino, A.; Smailagic, A.; Siewiorek, D.P. AHA-3D: A Labelled Dataset for Senior Fitness Exercise Recognition and Segmentation from 3D Skeletal Data. In Proceedings of the BMVC 2018, Newcastle, UK, 3–6 September 2018; p. 332.
93. Blanchard, N.; Skinner, K.; Kemp, A.; Scheirer, W.; Flynn, P. “Keep Me in Coach!”: A Computer Vision Perspective on Assessing ACL Injury Risk in Female Athletes. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1366–1374.
94. Parmar, P.; Morris, B.T. Learning to score olympic events. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 20–28.

95. Parmar, P.; Morris, B.T. Action Quality Assessment Across Multiple Actions. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1468–1476.
96. Parmar, P.; Morris, B.T. What and How Well You Performed? A Multitask Learning Approach to Action Quality Assessment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019, Long Beach, CA, USA, 15–21 June 2019; pp. 304–313.
97. Xu, C.; Fu, Y.; Zhang, B.; Chen, Z.; Jiang, Y.G.; Xue, X. Learning to Score Figure Skating Sport Videos. *IEEE Trans. Circuits Syst. Video Technol.* **2019**. [[CrossRef](#)]
98. Xiang, X.; Tian, Y.; Reiter, A.; Hager, G.D.; Tran, T.D. S3d: Stacking segmental p3d for action quality assessment. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 928–932.
99. Lea, C.; Flynn, M.D.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal convolutional networks for action segmentation and detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 156–165.
100. Qiu, Z.; Yao, T.; Mei, T. Learning spatiotemporal representation with pseudo-3d residual networks. In Proceedings of the IEEE International Conference on Computer Vision 2017, Venice, Italy, 22–29 October 2017; pp. 5533–5541.
101. Li, Y.; Chai, X.; Chen, X. End-to-end learning for action quality assessment. In Proceedings of the Pacific Rim Conference on Multimedia 2018, Hefei, China, 21–22 September 2018; pp. 125–134.
102. Li, Y.; Chai, X.; Chen, X. ScoringNet: Learning Key Fragment for Action Quality Assessment with Ranking Loss in Skilled Sports. In Proceedings of the Asian Conference on Computer Vision 2018, Perth, Australia, 2–6 December 2018; pp. 149–164.
103. McNally, W.; Vats, K.; Pinto, T.; Dulhanty, C.; McPhee, J.; Wong, A. GolfDB: A Video Database for Golf Swing Sequencing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2019, Long Beach, CA, USA, 15–21 June 2019.
104. Yadav, S.K.; Singh, A.; Gupta, A.; Raheja, J.L. Real-time Yoga recognition using deep learning. *Neural Comput. Appl.* **2019**, 1–13. [[CrossRef](#)]
105. Wang, Z.; Fey, A.M. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 1959–1970. [[CrossRef](#)]
106. Fawaz, H.I.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.A. Evaluating surgical skills from kinematic data using convolutional neural networks. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention 2018, Granada, Spain, 16–20 September 2018; pp. 214–221.
107. Funke, I.; Mees, S.T.; Weitz, J.; Speidel, S. Video-based surgical skill assessment using 3D convolutional neural networks. *Int. J. Comput. Assist. Radiol. Surg.* **2019**, *14*, 1217–1225. [[CrossRef](#)]
108. Doughty, H.; Damen, D.; Mayol-Cuevas, W. Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6057–6066.
109. Doughty, H.; Mayol-Cuevas, W.; Damen, D. The Pros and Cons: Rank-aware temporal attention for skill determination in long videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019.
110. Li, Z.; Huang, Y.; Cai, M.; Sato, Y. Manipulation-skill assessment from videos with spatial attention network. *arXiv* **2019**, arXiv:1901.02579.
111. SPHERE-Staircase 2014 Dataset. Available online: <https://data.bris.ac.uk/data/dataset/bgresiy3olk41nilo7k6xpkqf> (accessed on 23 July 2019).
112. SPHERE-Walking 2015 Dataset. Available online: http://cs.swansea.ac.uk/~{}csadeline/datasets/SPHERE-Walking2015_skeletons_only.zip (accessed on 23 July 2019).
113. SPHERE-SitStand 2015 Dataset. Available online: http://cs.swansea.ac.uk/~{}csadeline/datasets/SPHERE-SitStand2015_skeletons_only.zip (accessed on 23 July 2019).
114. UI-PRMD Dataset. Available online: <https://webpages.uidaho.edu/ui-prmd/> (accessed on 23 July 2019).
115. AHA-3D Dataset. Available online: <http://vislab.isr.ist.utl.pt/datasets/> (accessed on 23 July 2019).
116. Tao, L.; Elhamifar, E.; Khudanpur, S.; Hager, G.D.; Vidal, R. Sparse hidden markov models for surgical gesture classification and skill evaluation. In Proceedings of the International Conference on Information Processing in Computer-Assisted Interventions 2012, Pisa, Italy, 27 June 2012; pp. 167–177.

117. Forestier, G.; Petitjean, F.; Senin, P.; Despinoy, F.; Jannin, P. Discovering discriminative and interpretable patterns for surgical motion analysis. In *Conference on Artificial Intelligence in Medicine in Europe 2017*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 136–145.
118. Zia, A.; Essa, I. Automated surgical skill assessment in RMIS training. *Int. J. Comput. Assist. Radiol. Surg.* **2018**, *13*, 731–739. [[CrossRef](#)] [[PubMed](#)]
119. GolfDB Database. Available online: <https://github.com/wmcnally/GolfDB> (accessed on 23 July 2019).
120. Yoga Dataset. Available online: <https://archive.org/details/YogaVidCollected> (accessed on 23 July 2019).
121. JIGSAWS Dataset. Available online: <https://cirl.lcsr.jhu.edu/research/hmm/datasets/> (accessed on 23 July 2019).
122. Gao, Y.; Vedula, S.S.; Reiley, C.E.; Ahmidi, N.; Varadarajan, B.; Lin, H.C.; Tao, L.; Zappella, L.; Bejar, B.; Yuh, D.D.; et al. The JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling. In *Proceedings of the Modeling and Monitoring of Computer Assisted Interventions (M2CAI)—MICCAI Workshop*, Boston, MA, USA, 14–18 September 2014; p. 3.
123. Ahmidi, N.; Tao, L.; Sefati, S.; Gao, Y.; Lea, C.; Haro, B.B.; Zappella, L.; Khudanpur, S.; Vidal, R.; Hager, G.D. A Dataset and Benchmarks for Segmentation and Recognition of Gestures in Robotic Surgery. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 2025–2041. [[CrossRef](#)] [[PubMed](#)]
124. EPIC-Skills 2018 Dataset. Available online: <http://people.cs.bris.ac.uk/~{}damen/Skill/> (accessed on 23 July 2019).
125. BEST 2019 Dataset. Available online: <https://github.com/hazeld/rank-awareattention-network> (accessed on 23 July 2019).
126. The Breakfast Actions Dataset. Available online: <http://serre-lab.clps.brown.edu/resource/breakfast-actions-dataset/> (accessed on 23 July 2019).
127. ADL Dataset. Available online: <https://www.csee.umbc.edu/~{}hpirsiav/papers/ADLdataset/> (accessed on 23 July 2019).
128. Rohrbach, M.; Amin, S.; Andriluka, M.L.; Schiele, B. A Database for Fine Grained Activity Detection of Cooking Activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2012*, Providence, RI, USA, 16–21 June 2012; pp. 1194–1201.
129. Sigurdsson, G.A.; Varol, G.; Wang, X.; Farhadi, A.; Laptev, I.; Gupta, A. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, 11–14 October 2016; pp. 510–526.
130. Damen, D.; Doughty, H.; Maria Farinella, G. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV 2018)*, Munich, Germany, 8–14 September 2018; pp. 720–736.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).