*electronics*

MDPI

*Article*

# Learning Effective Skeletal Representations on RGB Video for Fine-Grained Human Action Quality Assessment

**Qing Lei [1,2,3], Hong-Bo Zhang [1,2,3], Ji-Xiang Du [1,2,3], Tsung-Chih Hsiao [1,3,*] and Chih-Cheng Chen [4,*]**

1 College of Computer Science and Technology, University of Huaqiao, Xiamen 361021, China; leiqing@hqu.edu.cn (Q.L.); zhanghongbo@hqu.edu.cn (H.-B.Z.); jxdu@hqu.edu.cn (J.-X.D.)
2 Fujian Key Laboratory of Big Data Intelligence and Security, Huaqiao University, Xiamen 361021, China
3 Xiamen Key Laboratory of Computer Vision and Pattern Recognition, University of Huaqiao, Xiamen 361021, China
4 School of Information Engineering, Jimei University, Xiamen 361021, China
* Correspondence: hsiaotc@hqu.edu.cn (T.-C.H.); 201761000018@jmu.edu.cn (C.-C.C.)

Received: 23 January 2020; Accepted: 26 March 2020; Published: 28 March 2020


**Abstract:** In this paper, we propose an integrated action classification and regression learning framework for the fine-grained human action quality assessment of RGB videos. On the basis of 2D skeleton data obtained per frame of RGB video sequences, we present an effective representation of joint trajectories to train action classifiers and a class-specific regression model for a fine-grained assessment of the quality of human actions. To manage the challenge of view changes due to camera motion, we develop a self-similarity feature descriptor extracted from joint trajectories and a joint displacement sequence to represent dynamic patterns of the movement and posture of the human body. To weigh the impact of joints for different action categories, a class-specific regression model is developed to obtain effective fine-grained assessment functions. In the testing stage, with the supervision of the action classifier's output, the regression model of a specific action category is selected to assess the quality of skeleton motion extracted from the action video. We take advantage of the discrimination of the action classifier and the viewpoint invariance of the self-similarity feature to boost the performance of the learning-based quality assessment method in a realistic scene. We evaluate our proposed method using diving and figure skating videos of the publicly available MIT Olympic Scoring dataset, and gymnastic vaulting videos of the recent benchmark University of Nevada Las Vegas (UNLV) Olympic Scoring dataset. The experimental results show that the proposed method achieved an improved performance, which is measured by the mean rank correlation coefficient between the predicted regression scores and the ground truths.

**Keywords:** action quality assessment; human activity analysis; skeletal feature representation

## 1. Introduction

Human action evaluation (HAE) aims to tackle the challenging problem of making computers automatically quantify how well people perform actions. It has been largely unexplored in past decades [1,2], and has been involved in a wide range of applications, such as sport activity scoring and training systems [1,3], physical therapy and rehabilitation [4–7], interactive entertainment [8–10], skill training for expertise learners, and video retrieval [11–13]. With the rapid progress in human activity understanding in the research area of computer vision, research efforts have recently been devoted to human action quality assessment [14,15].

Since the traditional manual assessment of human motion quality needs a great deal of expertise from specialized fields, longtime learning, and training processes are required to summarize the experience and evaluation rules for automatic scoring sport activity in a specialized field. This requires a great amount of time and high labor cost. Apart from traditional action recognition research, human action evaluation aims to design computation models for automatically assessing the quality score of human actions or activities and further give interpretable feedback to improve human body movement. It relies on accurate human motion detection and segmentation, action feature extraction and representation, and effective evaluation methods for measuring the quality of action performance. Severe challenges have to be dealt with when the action evaluation learning method is applied in realistic scenes such as intra-class variations in the scale, appearance, illumination, view, and inter-class ambiguity.

Most of the published research on human action evaluation directly employed advanced action recognition approaches to segment an action video into several action fragments, extract local or holistic motion features for video fragments, and aggregate fragmented features into a final feature representation. Then, regression models [14,15] or Hidden Markov Models (HMM) [16] were trained to estimate the quality score of featured actions. Furthermore, interpretable feedback was provided for improving the action performance. Some of these studies employed an optimization framework to formalize the action quality assessment problem [1,5,15]. They commonly employed this framework to develop a unified regression model for all action categories. One of the disadvantages of using a unified regression model is that large approximation errors caused by in-class variation lead to poor fitting in regression analysis. The second disadvantage is that a common evaluation function shared among all action categories can be fragile when dealing with unbalanced distributions of training data. Third, similar postures shared by different action categories significantly decrease the performance of action evaluation methods, such as the swing in a tennis serve and badminton smash, the spin in figure skating, and the floor exercise. Consequently, single assessment function learning for all action categories tends to generate an inaccurate assessment score and even provide the wrong feedback information. With significant progress in pose estimation methods, skeleton data of the human body can be estimated from an RGB video to facilitate detailed motion quality analysis. Most of the existing research introduced alignment and normalization methods developed for action recognition in action quality analysis to preprocess the skeleton data [17–20]. However, view variation due to a change of the camera position has not yet been addressed, despite the fact that it cannot be trivial for human action evaluation of a realistic dataset.

In this paper, we attempt to extract effective skeletal feature representation of human motion from an RGB video for a fine-grained human action quality assessment and develop a learning framework for assessing the action quality of sport activity. First, it is reasonable to assume that the action quality directly depends on the dynamic changes in human body movements. Therefore, in this study, we employed the OpenPose estimation method [21] to extract skeleton data from RGB videos for action quality analysis. Then, to determine the action category of test videos, we extracted the local motion pattern from each joint movement volume and aggregated all volumes to form a feature description for action classification. To accurately predict the quality score of an action video, we developed effective self-similarity feature descriptors extracted from the self-similarity matrices (SSMs) of joint trajectories and a joint displacement sequence that has been proven to alleviate the impact of camera motion in diving, figure skating, and vaulting videos. Lastly, a class-specific regression model learning strategy was employed to weigh the impact of joints on different action category evaluations. In the testing stage, with the supervision of the action classifier's output, the target regression model was determined to predict the quality score of the testing action video. The framework of our approach is illustrated in Figure 1.

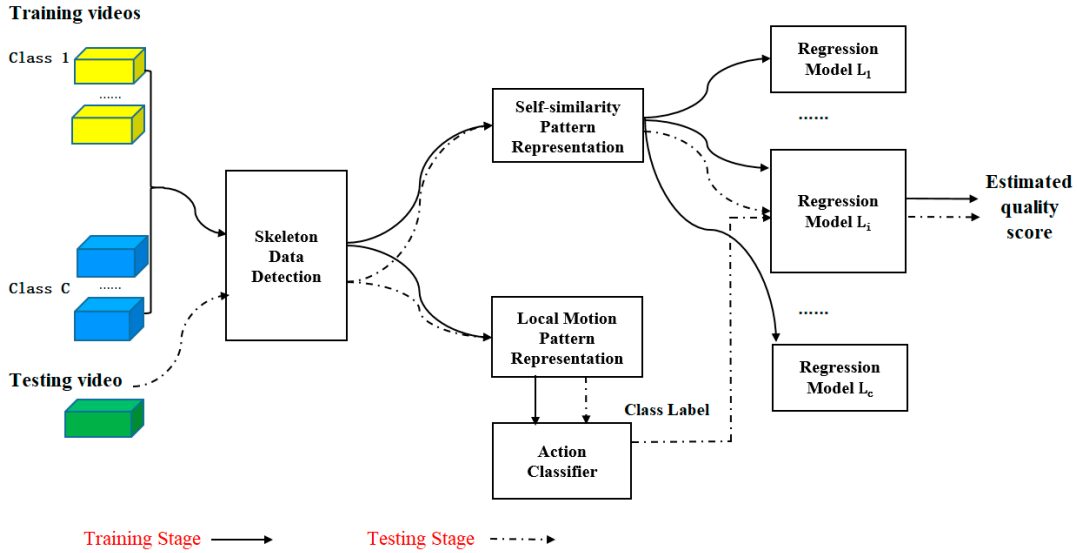



**Figure 1.** The framework of our action quality assessment method. (The input data were RGB videos of diverse action categories. Skeleton data detection: OpenPose algorithms provided by the work of Reference [21] were performed to capture skeleton data from the RGB video. Local motion pattern representation: the local motion pattern was extracted from each joint movement volume and aggregated to train the action classifier. Action classifier: the Support Vector Machine (SVM) classifier was developed from labeled training samples to determine the class label of the action video. Self-similarity pattern representation: we developed self-similarity feature descriptors extracted from joint trajectories and joint displacement sequences to represent the periodic property of sport activities. Regression models: we employed a class-specific learning strategy, and trained multiple regression models specific to different action categories for action quality score estimation.).

The contributions of this paper can be summarized as follows.

(1) First, in this paper, we propose an integrated action classification and action quality regression learning framework for the human action evaluation of RGB videos. An action classifier and assessment regression models are utilized, respectively, in different components. The former is used to predict the class label of a testing video, while the latter is employed to estimate the quality score with the supervision of a class label.

(2) Second, taking advantage of the view invariant property provided by self-similarity, in this paper, we develop self-similarity feature representation extracted from joint trajectories and joint displacement sequences to describe motion patterns of joints and posture changes, respectively. This encodes not only the dynamic changes of individual joints, but also the layout changes of all body joints. The experimental results prove that it alleviates the impact of camera motion in realistic scenes, and improves the skeleton representation's performance for diving, figure skating, and vaulting videos of an Olympic sports event.

(3) Experiments on a benchmark dataset were carried out in this study. The results show that the proposed method improved the rank correlation coefficient of the predicted scores against the judge's scores when compared with the baseline and other handcrafted feature methods.

The remainder of this paper is organized as follows. Section 2 introduces the related research works. Section 3 describes the algorithms developed to implement our action quality assessment method. After showing experiments in Section 4, we conclude the study in Section 5.

## 2. Related Works

On the basis of skeleton data analysis, several video-based human action evaluation research studies have been published in the last decade. There are two major issues focused on in these

studies. The first relates to capturing useful features from an RGB video to represent human actions for video-based human action evaluation, and the second relates to developing a robust evaluation method for accurately assessing the similarity between action features on account of complex environments. Some of the reviewed published works regarded the task of human action evaluation as a video sequence recognition problem. They tackled the challenge with the help of a machine learning method. In summary, the reviewed action quality assessment models proposed for human action evaluation research can be divided into three categories. These three categories include linear regression models (LR) [1,5,15], Hidden Markov Models (HMM) [4,7,10], and other statistic-based learning models [3,6,11,12,14].

In the early work of Reference [11], linear regression was employed to reduce the raw Motion Capture data for online action recognition. They designed a graph-based action model embodied with recurrent transitions for motion retrieval. The algorithm was linear and incremental, which makes it convenient for adding new actions and suitable for real-time application. Pirsiavash et al. [1] first proposed a two-layer processing framework for video-based human action quality assessment. In the first layer, they extracted spatio-temporal interest point features from regions returned by a human detection algorithm and computed discrete cosine transformation (DCT) features of joint displacement vectors from body joints' trajectories to represent human actions. Then, in the second layer, they developed a regression estimator to predict the quality score of a sports activity. Venkataraman et al. [15] tried to encode human actions through the dynamic changes of each body joint and the relationship between body joints. They developed two kinds of entropy-based features extracted from human skeleton data to represent these two clues and used them to realize human action segmentation from long video sequences and assess the quality score of diving in sports competition videos. Antunes et al. [5] presented a visual and human-interpretable feedback system for assisting with the physical activities of patients or athletes suffering from sport injuries. They also used skeleton data extracted from videos to quantify the action quality of human movements. First, the pre-processing transformation steps were conducted in both spatial and temporal dimensions to align a testing sequence with the template. Then, the matching error between the testing sequence and the template was computed based on the Euclidean distance of joints' coordinates in order to quantify the similarity between a testing sequence and a normal one. Furthermore, feedback for guiding how to perform properly was computed by minimizing the skeleton matching error and returned to the users.

Paiement et al. [4] used 3D skeleton data captured by two kinds of depth sensors and pre-processed the coordinates of joint sequences for online estimation of the quality of human movements. They proposed two statistical models: one was the probability density function (PDF) of each individual pose, and the other was the conditional PDF of a pose sequence in order to represent the features of normal movements. Then, log-likelihoods of observations compared with the model of normal ones were computed for quality assessment. They evaluated their methods using a gait on the stairs' dataset. In their further work of Reference [7], they studied four low-level pose features such as joint positions, joint velocities, pairwise joint distances, and pairwise joint angles. They also compared three kinds of discrete-state HMM and one continuous-state HMM to represent pose features and temporal dynamics of motion. They tested these features and models using periodic and non-periodic motions, including walking on a flat surface, gait on stairs, sitting, and standing. The experimental results demonstrated that continuous-state HMM performed better when describing motion dynamics for these action categories than other models for a frame-by-frame analysis of a motion quality assessment.

To concurrently evaluate the relevant spatial and temporal information of motion, Morel et al. [3] proposed an automatic morphology-independent and sport-independent method for assessing the motion of a player by comparing the features with the model learned from experts' motions. To deal with the different motion durations, they employed Dynamic Time Warping (DTW) to temporally align the skeleton data of joint trajectories. Considering the limitation of DTW—that the first and last frame of two aligned sequences are required to be in correspondence—Baptista et al. [6] investigated adapting Subsequence Dynamic Time Warping (SS-DTW) and Temporal Commonality Discovery

(TCD) to provide feedback proposals for improving the action performance of stroke survivors in a video-based rehabilitation system. Hu et al. [12] presented an action tutor system, which aimed to achieve a high-level evaluation of human action movements with the aid of Kinect. The body-joint configuration and shape/depth distribution of the human silhouette were encoded as pose descriptors to reflect the difference between various postures. Modified DTW and approximate string matching were further proposed to measure the similarity of actions. In the work of Reference [14], a novel framework for motion analysis, including the real-time action detection, recognition, and evaluation of motion capture data, was presented. The descriptor named Gesturelet was calculated for 3D skeleton joint positions, and combined the Moving Pose [22] and the angle descriptor [23] with appropriate weighting. Kinetic energy features were employed to construct Bag-of-Words data representation for action segmentation. In the evaluation component, 3D joint position and linear velocity errors were calculated and normalized, and then fed into a fuzzy logic engine to produce semantic feedback interpretations.

In this work, we investigated the effectiveness of using an integrated learning framework combining action quality assessment with action classification to boost the performance of action evaluation in realistic scenes. In this framework, a novel feature descriptor extracted from the self-similarity matrix of joint trajectories and joint displacement sequences was developed to alleviate the impact of camera motion. Then, class-specific regression models were established to predict the quality scores of action videos. Additionally, we trained the action classifier to supervise the determination of the regression model to assess the quality score of a testing action video. The experimental results proved that this approach is helpful for alleviating approximation errors caused by inter-class confusion.

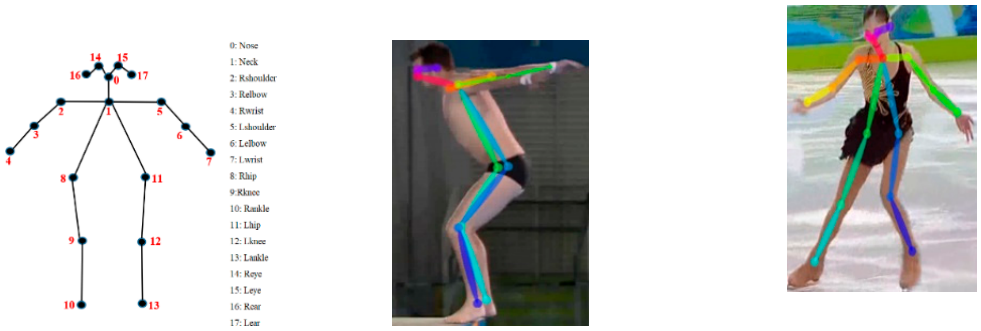
## 3. Proposed Method

### *3.1. Skeleton Data Extraction*

It is a reasonable assumption that the quality of human motion directly depends on the changing process of human body movement, which can be represented by the motion trajectory and relative location relationship between joints or body parts. Therefore, learning effective action features from motion trajectories of action videos plays an important role in developing reliable quality assessment algorithms for action evaluation. Most recent works employed skeleton data that was captured or detected from a depth or color camera for action evaluation research. With significant progress in recent pose estimation techniques and methodologies, skeleton data can now be estimated from RGB image data. Traditional skeletonization models, such as the deformable part model and flexible mixtures of parts model, have been replaced by deep neural network-based approaches. OpenPose [21] is an effective pose estimation method developed by the perceptual computing lab of Carnegie Mellon University. It is the first real-time multi-person skeleton detection system and works well when applied to RGB videos. Therefore, to obtain skeleton data of an action performer in an RGB video, we employed the state-of-the-art OpenPose algorithm to detect joints' position for each frame and extracted the trajectories of joints to represent the action video.

OpenPose provides the functionality of 2D real-time multi-person key point detection (15-, 18-, or 25-key point body/foot key point estimation). The 18-key point skeleton model is composed of 18 human body joints, as illustrated in Figure 2a, 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. Since action quality assessment is highly dependent on the changing positions and configuration of human body parts, the motion changes of key points on eyes, ears, and feet are not clear. Consequently, we only used the motion information of 14 body joints (N = 0~13) except for the eyes and ears for analysis. After the multi-person's skeleton detection results were returned by OpenPose, we carried out scale computation of the human body and a key point confidence comparison to extract the target performer's skeleton data and filter out noise data, which

results from a cluttered background. In the case of occlusion or self-occlusion, zero values of the joints' coordinate were obtained due to the failure detection of the human body. Then, linear interpolation was employed to capture the missing skeleton data from the pose estimation results of the previous frame and the next frame. In this way, the joint trajectories of the target performer were obtained. Some detection examples of diving and figure skating videos from the MIT Olympic Scoring Dataset [1] are illustrated in Figure 2b,c.




(a) (b) (c)

**Figure 2.** The 18-key point skeleton model and detection examples of OpenPose [21]. (**a**) Skeleton model of 18 joints, (**b**) example of diving, and (**c**) example of figure skating.

### *3.2. Action Classification Component*

We considered local feature representations' advantage in terms of their robustness in dealing with intra-class variation, and the fact that holistic feature representations provide a comprehensive description of human body movement for a time sequence. Combining both of their strengths, we proposed a joint movement feature to acquire discriminative information for action classification. It is believed that the semantics of human action are related to the movement pattern of joints and the relationship of the human body interacting with its surrounding environment. Therefore, building effective features that capture both the discriminative dynamics of body joints and the descriptive spatial context for class label determination is investigated in this paper. The pipeline of our action classification component is illustrated in Figure 3.

Let $\left\{S^1, S^2, \ldots, S^N\right\}$ denote a set of joint trajectories obtained by skeleton detection for video V, where N is the number of human body joints, $S^k = \left[s_1^k, s_2^k, \ldots, s_T^k\right]$ represents the trajectory of the kth joint, and T is the number of frames of V. Each joint position is located by its coordinates $\mathrm{s} = \left[s_x, s_y, s_t\right]$ in discrete (x, y, t)-space. To capture the spatial context of joints, a n × n dimensional local patch centered at each joint position $\mathrm{s}^k = \left[s_x{}^k, s_y{}^k, s_t{}^k\right]$ is extracted from video frames where $k = 1 \sim N$. All patches extracted over a temporal duration ($t = 1 \sim T$) are assembled into a motion volume of joint k, denoted by $v^k (k = 1 \sim N)$, which is illustrated by the red cuboids in Figure 3.

To filter out noise data resulting from failure or false detection of the joint position, 2D Gaussian smoothing is first performed for the joint motion volume $v^k$. Then, the central moment features $m_{i,j}^r (\mathrm{i,j} = 1 \sim \mathrm{n}, \mathrm{r} = 1,2)$ for each super-pixel $v_{i,j}^k$ of joint motion volume $v^k$ are calculated according to Equation (1).

$$m_{i,j}^r = \frac{1}{T}\Sigma_{t=1}^{T}\left(G_{i,j,t} - \overline{G}\right)^r, \ \overline{G} = \frac{1}{n \times n \times T}\Sigma_{t=1}^{T}\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}G_{i,j,t} \quad (1)$$

where $G_{i,j,t}$ is the value of the pixel located at the (i, j, t)-coordinate in filtered volume $v^k$. Features of all the super pixels contained in a joint motion volume $v^k$ are assembled to form the motion feature $m^k$, and to represent the motion pattern of the kth joint. Then, all the features of N joint volumes are concatenated to form the final spatio-temporal feature description to represent the action instance occurring in video V.

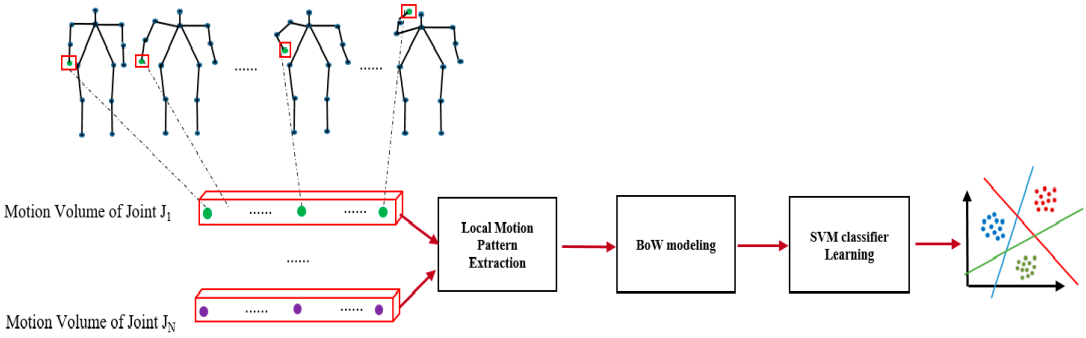



**Figure 3.** The pipeline of the action classification component. (The input of the action classification component was joint trajectories obtained by skeleton detection from an RGB video. Joint motion volume: A joint motion volume centered at each of the joint positions and comprised of the $n \times n$ spatial context was extracted to represent the movement of each joint. Local motion pattern representation: We employed the central moment features computed by the noise filtered joint motion volume to describe the motion patterns of each corresponding joint, and concatenated all joint features into the final motion descriptor. Bag-of-words (BoW) modeling: For the sake of alleviating intra-class variation, we employed bag-of-words feature modeling, which utilizes K-means feature clustering and cluster frequency statistics to form the final action representation of a video. Action classifier: The SVM classifier was developed from features of labeled training samples to determine the class label of action instance.).

We propose applying the Bag-of-Words (BoW) [24] model for action representation. Specifically, the unsupervised K-means algorithm is first performed on all joint motion volume features extracted from training videos to obtain K clusters for constructing the action codebook. The center of each cluster is called a visual word, and all centers of K clusters form a visual codebook for action modeling. Then, the original feature is projected onto the closest visual word in the action codebook. All features of an action video are projected and aggregated into an occurrence frequency histogram of visual words. This forms the final BoW feature representation of the action video.

Lastly, BoW features with class labels are fed into maximum margin classifier learning, as formulated in Equation (2).

$$min_{w,b} \frac{\|w\|^2}{2}, s.t.\ Y\left(w^T H_C + b\right) - 1 \geq 0, \tag{2}$$

where $H_C$ denotes the feature vector of a training video for classification and Y is the ground-truth action class label of each training video provided by manual annotation in the benchmark dataset. w and b represent the normal vector and bias of the classification hyperplane, respectively.

### *3.3. Quality Assessment Component*

As stated above, skeleton-based pose feature representations intuitively reflect the changing process of human body movement and provide significant information for action quality assessment. Therefore, when developing an accurate action quality assessment method, it is preferable to encode the dynamic changes of joints or body parts into feature representation for action analysis. However, the fine-grained quality assessment of human action is confronted by the challenges of intra-class variations, such as different scales of the human body, variation in motion velocity, individual style, and changes due to camera motion. We performed skeleton data preprocessing, including noise filtering, scale normalization, and spatial alignment, to tackle the problem of intra-class variations. Furthermore, self-similarity patterns were extracted from joint trajectories and joint displacement sequences, respectively, to model the invariant dynamics of human body motion and to deal with view changes in realistic scenes.

Another motivation is the observation that the movement ranges of joints are different for various action categories. Therefore, the importance of different joint movements cannot be considered

identical for action quality evaluation. For example, as far as serves, smashes, and swings in tennis and badminton activity are concerned, the movement of upper limbs ranges more significantly than that of lower limbs. On the contrary, the remarkable change that occurs in lower limb motion should be addressed in pommel horse riding, parallel bars, and football. Furthermore, all limb movement may be considered equal for figure skating and diving quality assessments. Therefore, it should be noted that, for different action categories, the similarity measurement of quality features should address weight assignment for the impact of different joints. To address this problem, a class-specific regression model was developed to weigh different impacts of joint movements and to obtain more accurate evaluation scores for fine-grained human action quality assessment in this work.

The quality assessment component of our proposed learning framework consists of the preprocessing of joint trajectories, joint displacement sequence extraction, feature extraction from the self-similarity matrix (SSM) of joint trajectories and joint displacement sequence, and class-specific regression learning. The training process of our class-specific action quality assessment component is illustrated in Figure 4.

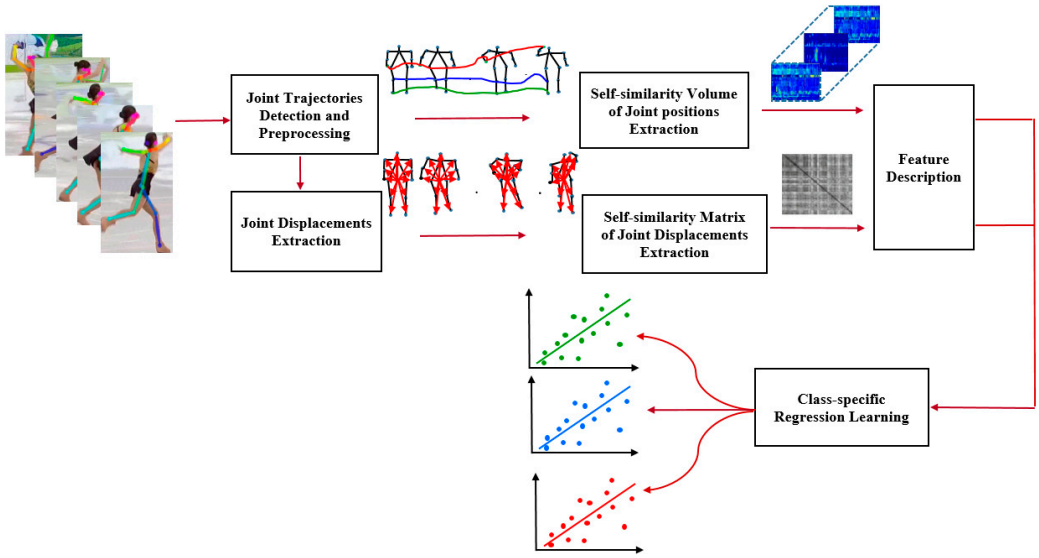



**Figure 4.** The training process of our quality assessment component. (The input of the action classification component was joint trajectories obtained by skeleton detection from an RGB video. Joint trajectory detection and preprocessing: We performed noise filtering, normalization, and spatial alignment processes for detected joint trajectories to deal with failure or false detection, scale variation, and spatial transformation of human motion. Joint displacement extraction: As the middle of the left hip and right hip was selected as the central point, the displacement of each joint relative to the central point was computed to represent the layout relationship between human body parts. Self-similarity volume of joint position extraction: Temporal self-similarity matrix of each joint trajectory was computed to capture view invariant patterns of intra-joint dynamic changes. Self-similarity matrix of joint displacement extraction: Temporal self-similarity matrix of all joint displacements was computed to capture view invariant patterns of inter-joint dynamic changes. Feature description: the Histogram of Gradient (HOG) descriptor was employed to describe the pattern structure of two kinds of self-similarity matrices. Class-specific regression learning: We utilized two types of regression strategy—support vector regression and ridge regression—to develop class-specific regression models for an action quality assessment.).

### 3.3.1. Pre-Processing of the Joint Trajectory

The original skeleton data detected from RGB videos often contain noise in cases of occlusion and cluttered environments in realistic applications. To obtain robust similarity quantization for fine-grained assessment, the first processing step of skeleton data is the noise filtering of incorrect joint

coordinates resulting from incorrect human detection. Then, normalization and alignment processing, including the scale, transition, rotation, and appearances, are subsequently conducted to deal with various intra-class variations.

In the case of occlusion or self-occlusion, the failure or false detection of the human body leads to outliers of joint coordinates in skeleton estimation that lead to an unreliable representation of human motions. In this study, discrete cosine transformation was performed for discrete coordinates of each joint trajectory to filter out zero values or sharply changing coordinates resulting from failure or false pose estimation. Low-frequency components were preserved for the reliable detection of human body positions.

On one hand, the scale of captured skeleton data can be quite diverse due to different distances between the subject and camera. The original joint coordinates are required to be normalized to a prototypical range for comparison. The scale normalization process is formulized as follows. Suppose that you are given skeleton data $\left\{S^1, S^2, \ldots, S^N\right\}$ detected from an action video I, where N denotes the number of joints and $S^k = \left[s_1^k, s_2^k, \ldots, s_T^k\right]$ represents the motion trajectory of the $k$th joint through T consecutive frames of I. First, the middle of Rhip (Joint 8) and Lhip (Joint 11) of the first frame is defined as the central point, and the distance from Neck (Joint 1) to the central point is defined as the normalized length. Then, the joints' coordinates are scaled by the normalized length. The normalization process can be formulated by the following equation.

$$s'^k_t = \frac{s_t^k}{\| s_1^1 - \frac{s_1^8 + s_1^{11}}{2} \|}, \tag{3}$$

where $s_t^k = \left[s_x, s_y, t\right]$ denotes the origin coordinate of joint k in frame t in the (x, y, t)-space.

Then, spatial alignment is conducted by performing rotation transformation. The rotation angle $\theta$ is determined by projecting the vector from Lhip to Rhip onto the $x$-axis. Then, each joint's coordinate is transformed by rotating θ-degree, as formulated by Equation (4).

$$\theta = \text{acos}\left(\frac{\overrightarrow{\left(s^{11} - s^8\right)} \cdot \vec{ox}}{\left\|s^{11} - s^8\right\| \| \vec{ox} \|}\right), \quad \vec{s}' = \begin{bmatrix} cos\theta & sin\theta \\ cos\theta & -sin\theta \end{bmatrix} \vec{s}, \tag{4}$$

where, $\vec{s} = \left[s_x, s_y\right]$ and $\vec{ox} = [1, 0]$.

### 3.3.2. Self-Similarity Feature Description

The motion pattern for different viewpoints varies significantly for an action, as shown in Figure 5. It illustrates two different joint trajectories of the same person performing a diving action in the side and front views. The first row presents the skeleton detection results on the side view and the second row shows the skeleton detection results on the front view. It should be noted that a reliable feature representation for fine-grained quality assessment has to be robust for different camera positions.

Junejo et al. [25] introduced the use of trajectory-based self-similarity matrices (SSMs) to encode humans observed from different views for classifying human actions. They proved that the Histogram of Gradient (HOG) and Optical Flow (OF) features extracted from a self-similarity matrix are stable under view changes of an action. In Reference [25], for calculating the trajectory-based SSM of $\left\{S^1, S^2, \ldots, S^N\right\}$, each element of SSM is computed by the equation below.

$$d_{ij} = \frac{1}{N} \sum_{k=1}^{N} \| s_i^k - s_j^k \|_2, \tag{5}$$

where $s_t^k = \left[s_x, s_y, t\right]$ denotes the original coordinate of joint k in frame t in the (x, y, t)-space, and $\| \cdot \|_2$ represents the Euclidean distance.

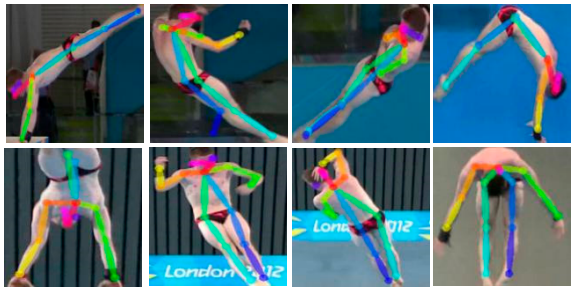


**Figure 5.** Joint trajectories of two different views for the same person performing diving. (The first row shows results on the side view and the second row presents results on the front view.).

It is worth noting that each element of the temporal self-similarity matrix in Reference [25] represents the accumulated coordinate's offset over all body joints for a frame of $t$, as formulated in Equation (5). However, it completely neglects the individual motion dynamics of each joint and the relationship between body joints' relative positions. We believe that both of these factors can be essential for an action similarity measurement and should be addressed in feature representation. Different from the SSM feature in Reference [25], we represent the trajectories of each joint of the human body independently, and further analyze the displacement sequence of inter-joints to build the temporal self-similarity matrices. As a result, temporal self-similarity matrices of joint trajectories and displacement sequences are computed, respectively, to capture view invariant patterns of intra-joint and inter-joint dynamics.

The calculation process is formulized as follows. Suppose that you are given preprocessed skeleton data $\left\{S^1, S^2, \ldots, S^N\right\}$ through the procedures outlined in Section 3.3.1, where N denotes the number of joints and $S^k = \left[s_1^k, s_2^k, \ldots, s_T^k\right]$ represents the motion trajectory of the $k$th joint through $T$ frames of the image sequence. The temporal self-similarity matrices of joint trajectories can be denoted by $SSM_{J_S} = \left[SSM_{J_1}, SSM_{J_2}, \ldots, SSM_{J_N}\right]$, where $SSM_{J_k} = [d_{ij}^k]_{T\times T}$ and $d_{ij}^k$ denotes the Euclidean distance between the position coordinates of the $k$th joint in frame $i$ and $j$, as formulated by Equation (6).

$$d_{ij}^k = \| s_i^k - s_j^k \|_2 \tag{6}$$

The middle of Rhip (Joint 8) and Lhip (Joint 11) in each frame is regarded as the central point, and the displacement sequence of skeleton data can be represented by $\mathrm{P} = [p_1, p_2, \ldots, p_T]$, where $\vec{p_t} = \left[p_t^1, p_t^2, \ldots, p_t^N\right]$ and $p_t^k$ denotes the displacement of the $k$th joint relative to the central point in frame $t$. It is computed by the equation below.

$$p_t^k = |s_t^k - \frac{s_t^8 + s_t^{11}}{2}|, \tag{7}$$

where $|\cdot|$ denotes the norm of the vector. Then, the temporal self-similarity matrix of the joint displacement sequence can be denoted by $SSM_D = [d'_{ij}]_{T\times T}$, where $d'_{ij}$ denotes the Euclidean distance between two displacement vectors $\vec{p_i}$ and $\vec{p_j}$ belonging to frame i and j. It is computed by Equation (8).

$$d'_{ij} = \| \vec{p_i} - \vec{p_j} \|_2 = \frac{1}{N} \sum_{k=1}^{N} \| p_i^k - p_j^k \|_2 \tag{8}$$

The Histogram of Gradient (HOG) feature descriptor extracted from a log-polar semicircle centered on the main diagonal of the matrix is used to describe the pattern structure of the self-similarity

matrix of joint trajectories and joint displacement sequences. For a given semicircle area, $C = \{r_i, \theta_i\}(r_i = 0 \sim r, \theta_i = 0 \sim \pi)$, where r is the radius of the semicircle, which denotes the temporal window extent to be considered, and the origin of the pole coordinate is located at the *j*th element of SSM's main diagonal. C is segmented into 11 blocks equally divided into three sections of the polar axis and five bins of the polar angle. One block near the origin of the pole ordinate is kept within one-third of the radius and without division of the polar angle. For each block of C, the normalized eight-bin HOG feature vector $h_j^a = [h_{j,b}^a]'_{b=1\sim 8}$ is calculated, and the vectors of all 11 blocks are concatenated to form a feature vector $h_j = [h_j^a]'_{a=1\sim 11}$ of the jth element of SSM's main diagonal. $h = (h_1, h_2, \ldots, h_T)$ is computed for all elements of SSM's main diagonal to form a feature description of SSM. Lastly, HOG feature vectors of all joint trajectories and joint displacement sequences are concatenated to generate self-similarity feature representation of an action instance $H = [h_{SSM_{J1}}, \ldots, h_{SSM_{J_N}}, h_{SSM_D}]$.

#### 3.3.3. Class-Specific Regression Model

Not all joints are equally involved in the motion of the human body, and, for different action categories, the participating joints show different levels of significance. Therefore, different impact weights should be assigned to distinguish the importance of body joints. We propose a class-specific regression learning strategy to address this problem. In this strategy, training samples annotated with class labels and quality scores were used to train regression weight vectors for specific action categories, which resulted in a more effective evaluation of different categories that shared postures. This alleviated the confusion. Correspondingly, a more accurate score can be estimated to boost the performance of a fine-grained quality assessment. The learning process is formulated as follows.

The training set $D = \left\{\left(H_1^1, \mathrm{y}_1^1\right), \left(H_2^1, \mathrm{y}_2^1\right), \ldots, \left(H_{n_1}^1, \mathrm{y}_{n_1}^1\right), \ldots, \left(H_1^c, \mathrm{y}_1^c\right), \left(H_2^c, \mathrm{y}_2^c\right), \ldots, \left(H_{n_c}^c, \mathrm{y}_{n_c}^c\right)\right\}$ consists of action videos annotated with action class labels and quality assessment scores. $H_j^i \in R^K$ is the self-similarity feature vector of the jth action instance that belongs to action class *i*, which is computed through the feature description procedure described in the previous subsection. $y_j^i \in R$ denotes the ground-truth quality score and $n_i$ is the number of training samples belonging to action class *i*. The regression model $w^i$ of the *i*th action category is trained for all videos of the same action category by finding a real-valued linear function ${w^i}^T H_j^i$, which acquires the minimized approximation errors between the ground truths and estimated scores. The estimated score is obtained by projecting the original features onto the transformation space. The learning process is formulated as the following.

$$\mathrm{argmin}_{w^i} \sum_{j=1}^{n_i} \| \, y_j^i - {w^i}^T H_j^i \, \|_2, \tag{9}$$

where $w^i \in R^K$, $y_j^i$ is the ground-truth quality score of the *j*th action instance belonging to class *i* and ${w^i}^T H_j^i$ is the corresponding predicted score.

The algorithm is performed individually for each action category to obtain a specific regression model. In the proposed method, two types of regression strategies—Support Vector Regression (SVR) and Ridge Regression (RR)—are employed to implement this training process. During testing, with the supervision of the action classifier's output, the specific regression model is determined to predict the quality score of the testing video.

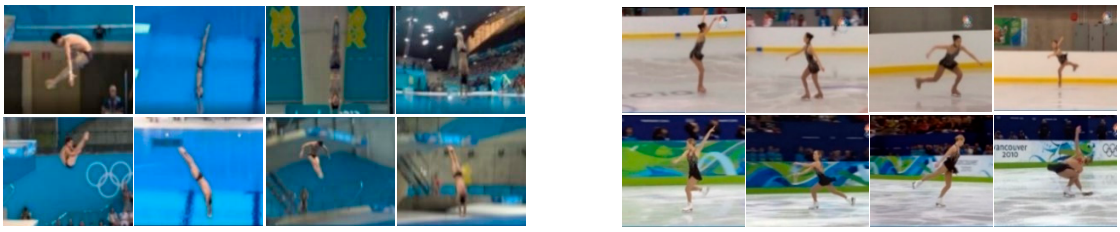
## 4. Experiments

The construction of an action evaluation dataset requires professional experts to annotate training videos according to their knowledge and experience of relevant fields. The annotation is highly dependent on the subjective judgement of professionals. It is difficult to collect training samples from massive action categories, and accurate annotation acquisition requires a great deal of manual effort when constructing a large-scale dataset of human action evaluation. Therefore, several limitations are shared with the published dataset of action quality assessment, such as insufficient training data, a limited number of action categories, a fixed position of the RGB camera, and identical scenes. In this

study, we investigated the performance of our proposed method for the diving and figure skating videos of the MIT Olympic Scoring dataset [26], and the gymnastic vaulting videos of the UNLV Olympic Scoring dataset [27,28].

### *4.1. Introduction of Datasets and Experimental Setups*

The MIT Olympic Scoring dataset [26] contains two action category sport videos: diving and figure skating. The diving dataset contains 159 videos that include 25,000 frames with scores varying between 20 and 100. The frame rate of diving videos is 60 fps and each diving instance is about 150 frames. The figure skating dataset consists of 150 videos captured at 24 fps. There is a total of 630,000 frames, and each action instance is almost 4200 frames. The ground-truth score of each action video is freely obtained by extracting the judge's score that is released publicly in sports footage. The quality assessment score of each action instance varies between 0 and 100. The figure skating assessment is more suitable for activity evaluation since several action components, such as jumps, spins, and steps, are included and repeated in each of the videos. Compulsory routines are required in the performances of the Olympic games, including a spin, axel, spiral, and transition. However, the sequential order of routines is different in action videos. Therefore, figure skating is more challenging than diving because of the complexity of activity analysis rather than simple actions. Each video is annotated with the start frame, end frame, and judgement score of action occurrence. The score is extracted from the frame by displaying the referee's decision from the video. Figure 6a,b show some examples of diving and figure skating frames performed by different people with different views.



(a) (b)

**Figure 6.** Example frames of the MIT Olympic Scoring dataset [26]. (**a**) Example frames of diving. (**b**) Example frames of figure skating.

The UNLV Olympic Scoring dataset [27,28] includes sport videos from Summer and Winter Olympics events on YouTube. It comprises 1189 videos from seven action categories, including 370 videos of single-person diving from a ten-meter platform, 88 videos of synchronized diving from a three-meter springboard, 91 videos of synchronized diving from a ten-meter platform, 176 videos of a gymnastic vault, 175 videos of a skiing sport, 206 videos of snowboarding, and 83 videos of trampolining in Olympic events. In the action categories of vaulting, skiing, and snowboarding, severe view changes existed in the action videos. Since the motion patterns of a single person's action are addressed in our feature representation, synchronized events of more than one person are not taken into account. Additionally, a long-range shot distance is commonly adopted in filming videos of skiing, snowboarding, and trampolining. The accuracy of skeleton detection is severely reduced with the influence of poor pose estimation results. Therefore, we only evaluated our proposed method for vault videos of this dataset. The vault dataset of UNLV Olympic Scoring includes 176 videos with an average length of about 75 frames. The frame size is $320 \times 240$. The ground-truth score of a vaulting video that ranges from 12.30 to 16.87 is determined by the sum of the "execution" score and "difficulty" score.

In the action classification component, we propose extracting local motion patterns from the joint motion volume and training the linear kernel-based SVM classifiers to determine the action label of the testing action video. The estimation action class of the testing video can be obtained by comparing the output of all classifiers and finding the highest category confidence. For the evaluation of our classification component, we adopted leave-one-video-out validation for 159 diving videos, 150 figure

skating videos, and 176 vaulting videos. The performance was measured by the average accuracy over all videos. The proposed classification method achieved the performance of 92.6% for all action videos. The average classification accuracies of diving, figure skating, and vaulting were 93%, 89.3%, and 94.8%, respectively.

For validating the proposed action quality assessment method, the rank correlation coefficient between the predicted scores and the ground truths was computed to evaluate the performance. We employed the rank correlation coefficient measurement of two vectors' similarity to evaluate our proposed method. It was computed by Equation (10).

$$\rho_{y,\hat{y}} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \tag{10}$$

where y is the ground truth score vector and $\hat{y}$ is the predicted score vector. $d_i$ denotes the ranking difference between $y_i$ and $\hat{y}_i$ of the $i$th video. A higher ρ value denotes a better estimation performance. In experiments using the MIT diving and figure skating dataset, we followed the testing schema introduced in Reference [1]. A random split of action videos was adopted, resulting in 100 videos being selected as training samples and the rest as test samples. The average rank correlation coefficient of 200 rounds' testing results was computed as the final evaluation performance of the dataset. Additionally, in experiments using the UNLV vaulting dataset, the testing schema introduced in Reference [27] employing a fixed split of the dataset, namely 120 videos for training and the remaining 56 videos for testing, was followed. The detailed results are presented in the next section.

### *4.2. Results of the MIT Diving Dataset*

According to the evaluation protocol introduced in Reference [1], 100 labeled action videos were selected randomly as training samples from the diving dataset and the remaining 59 videos were used as test samples. The rank correlation coefficient between the predicted scores and the ground truths of this round split was computed. Then, 200 rounds of experiments with different random splits were conducted according to the same strategy. The evaluation performance was obtained by averaging all rounds' rank correlation coefficients. We compared the performance of our proposed method with four state-of-the-art action feature methods for human action recognition, and several evaluation learning methods, including the all-action regression model, single-action regression models, support vector regression of different kernels, and ridge regression.

We compared three feature extraction methods, including spatio-temporal interest points (STIP) [29], dense sampling (Dense) [30], and skeleton data (Skeleton). The benchmark MIT Olympic Scoring dataset that we employed to evaluate the proposed method was published in Reference [1]. Additionally, a similar solution strategy was employed both in Reference [1] and our proposed learning framework. A handcrafted feature engine was first built for feature representation and a regression model was then developed from the action features for quality assessment. Therefore, we chose Reference [1] as the baseline method. In Reference [1], the researchers extracted both low-level spatial and temporal filtering features that captured edges and velocities, as well as high-level pose features obtained from the discrete cosine transformation of joint displacement vectors, and estimated a regression model that predicted the scores of actions. They compared their performance with the space-time interest points (STIP) method [29] and Discrete Fourier Transform (DFT) pose features. STIP is the abbreviation for space-time interest points and was developed for feature detection in traditional action recognition research. The method was presented based on the observation that actions frequently occurred in positions with sharp changes in both the spatial and temporal domains. It employed a space-time extension of the Harris corner detector to extract the prominent positions of significant changes in spatial and temporal dimensions from the action video. Then, histograms of oriented gradients (HOG) and the optical flow (HOF) were calculated and concatenated for each local spatial and temporal volume centered at the prominent position. All the local features were aggregated

to form the feature descriptor for action representation. However, it has been proven in Reference [30] that dense sampling has demonstrated a better classification performance than original STIP for human action recognition in realistic scenes. Dense sampling extracted video blocks at regular intervals and scales in space and time by a sliding window moving throughout the whole video. The HOG and HOF descriptors of each video block were computed and concatenated to represent the local feature of each spatial and temporal position. All local feature descriptors were aggregated to form the final feature representation of the whole video. Skeleton data can be obtained from RGB videos using the pose estimation algorithm. Pose features extracted by transformation and projection of the original skeleton data were regarded as being encoded with high-level semantics of human actions. Therefore, we compared the performance of our skeleton data extraction method with that of STIP [29] and dense sampling [30] using the benchmark MIT Olympic Scoring dataset.

On the other hand, we developed self-similarity feature representation extracted from joint trajectories and joint displacement sequences to describe motion patterns of joints and posture changes. The self-similarity matrices employed to encode human actions were initially proposed in Reference [25], in which only the coordinate's offset over all body joints was accumulated for a single frame, and the individual motion dynamics of each joint and the relationship of body joints' relative positions were neglected. In contrast to Reference [25], we encoded the motion dynamics of each body joint independently, as well as the displacement sequence between body joints, to build temporal self-similarity matrices. On the basis of skeleton data representation for human actions, we compared our proposed feature method with the baseline self-similarity matrix feature [25] and the pose feature captured from discrete cosine transformation (DCT) [1] for the original coordinates of human body joints. The performance comparison of our proposed feature and the reviewed features of the MIT diving dataset is presented in Table 1.

**Table 1.** Quality assessment results on the Massachusetts Institute of Technology (MIT) diving dataset. The average rank correlation coefficients between the predicted results and the ground truth of different feature representation and evaluation methods are presented (The higher the value, the better the performance.).

| | STIP * | Dense | Skeleton + DCT | Skeleton + SSM * | Our Method |
|---|---|---|---|---|---|
| **All-action SVR-Linear** | 0.07 | 0.09 | 0.19 | 0.13 | **0.20** |
| **Single-action SVR-Linear *** | 0.18 | 0.16 | 0.45 | 0.35 | **0.52** |
| **Single-action Ridge Regression** | 0.07 | 0.10 | 0.32 | 0.30 | **0.4** [1] |

[1] The bolded value indicated the best performance of all compared feature methods. * STIP means space-time interest points feature method. SSM indicates self-similarity matrices feature method. SVR means support vector regression method.

As shown in Table 1, skeleton-based feature representations exhibited a significantly greater strength than low-level features of STIP or dense sampling under different evaluation strategies for the diving video assessments. On the basis of the same skeleton data detection results, the DCT feature method captured from the discrete cosine transform on the original coordinates of joints achieved better performances at 0.19, 0.45, and 0.32 under different evaluation strategies, in comparison to the corresponding results of 0.13, 0.35, and 0.30 obtained by the SSM feature method. This is likely because the SSM feature accumulated the position offsets of all joints between each frame of the video sequences. Therefore, it completely discarded the motion information of individual joints of the human body and the relationship between body joints. The DCT feature captured the dynamic changes of human body movement. However, it simply considered the change of joints' position relative to the centroid of the human body along the time dimension, and neglected the motion patterns of each joint trajectory and the changes of the joints' correlation. The proposed feature method extracted patterns from the self-similarity matrix of each joint trajectory and joint displacement sequence. It encoded not only the dynamic changes of individual joints, but also the pose feature described by the layout

changes of all body joints. It achieved the best rank correlation coefficients of 0.20, 0.52, and 0.4 among all the experienced feature methods for the MIT diving dataset.

Furthermore, we compared the performance of different evaluation methods, namely all-action support vector regression (all-action SVR) evaluation, single-action support vector regression (single-action SVR) evaluation, and single-action evaluation ridge regression (single-action ridge regression) evaluation for this dataset. All-action evaluation trained a unified regression function to assess all action categories, and single-action evaluation employed the strategy of specific-action training to learn a specific assessment function for each action category from annotated action features. As shown in Table 1, the performance difference is not clear between single-action ridge regression and all-action evaluation under low-level STIP (single-action, 0.18, all-action, 0.07) and dense sampling features (single-action, 0.16, all-action, 0.09). However, the estimation results were significantly improved under the same skeleton-based feature representation, and the performance of single-action evaluation was significantly superior to that of all-action evaluation. This indicated that a dedicated quality assessment model for specific action categories is suitable for sport activity scoring. Therefore, it is less effective to design one unified feature evaluation function to assess various kinds of feature patterns' similarities for all action categories. We also attempted different kernel functions of support vector regression evaluation and compared the evaluation methods of SVR and Ridge Regression (RR). It was found that SVR with a linear kernel achieved a better performance of 0.52 than the ridge regression method, which displayed a value of 0.4. In addition, the linear kernel always achieved better results than the Radial Basis Function (RBF) kernel and sigmoid kernel in SVR regression.

The predicted score of each testing video for the MIT diving dataset from our best rank correlation coefficient performance is presented in Figure 7. The horizontal axis denotes the index of the action video for testing, and the vertical axis represents the predicted quality score. The data series of the GT_test indicate the ground truth scores of test videos, and the pred test corresponds to the estimated scores of the proposed method.

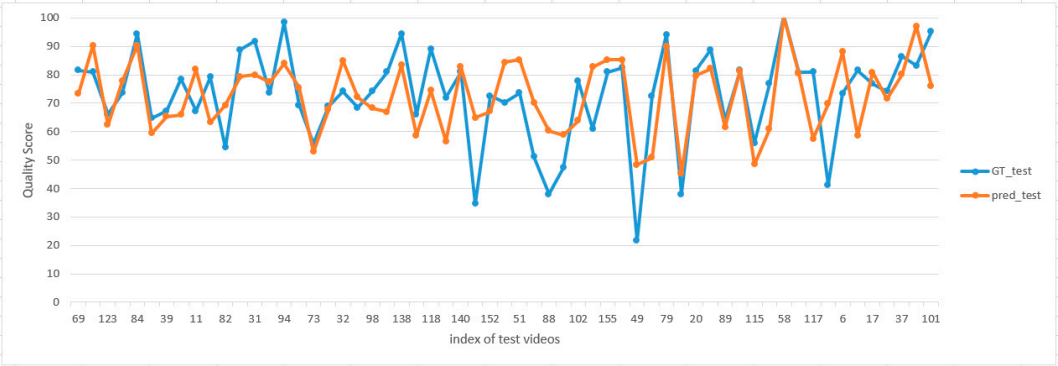



**Figure 7.** The predicted scores of test videos from our best rank correlation coefficient for the MIT diving dataset.

### *4.3. Results on the MIT Figure Skating Dataset*

The figure skating activity of the MIT Olympic Scoring dataset is more challenging than the diving action. The frame rate of figure skating videos is 24 fps and the length of each video is about 4200 frames. Therefore, complex activity, rather than a single action, is involved in a figure skating video. Other challenges are the irregular camera motion and cluttered background that commonly exist in long-duration recorded videos. Compulsory routines are required and performed successively in each video, such as jumps, spins, turns, steps, and moves. However, the order of execution for custom actions is not strictly restricted. Moreover, significant pose variations seriously affected the performance of action quality assessment due to the appearance of athletes, irregular changing of the camera position, and zooming in and out for the sake of capturing the best recording effect.

To obtain effective feature representation of complex activities in long-duration videos, we divided each video into 10 segments with an equal length. Feature extraction was performed for each segment, and all segments' features were concatenated to represent the action feature of the whole video. The performance using the figure skating dataset was also evaluated according to the testing protocol in Reference [1] where the average rank correlation coefficient of 200 experiments was computed with random splitting. In this case, 100 action videos were selected as training samples and the rest were selected as tests. The performance comparison of our feature method and the baseline method and other state-of-the-art feature methods was investigated. The details are presented in Table 2.

**Table 2.** Quality assessment results on the MIT figure skating dataset. The average rank correlation coefficients between the predicted results and the ground truth of different feature representations and evaluation methods are presented.

| | STIP * | Dense | Skeleton + DCT | Skeleton + SSM * | Our Method |
|---|---|---|---|---|---|
| **All-action SVR-Linear *** | 0.13 | 0.15 | 0.21 | 0.15 | **0.25** |
| **Single-action SVR-Linear** | 0.21 | 0.23 | 0.37 | 0.19 | **0.41** |
| **Single-action Ridge Regression** | 0.20 | 0.21 | 0.25 | 0.17 | **0.28** |

* STIP means space-time interest points feature method. SSM indicates self-similarity matrices feature method. SVR means support vector regression method.

From Table 2, it can be observed that most of the skeleton-based feature representations achieved a superior performance in comparison to low-level STIP or dense sampling feature methods. In skeleton data-based feature representation, the best mean rank correlation coefficients of the Discrete Cosine Transform (DCT) transformation feature and our proposed feature were 0.37 and 0.41, respectively, which outperformed STIP (0.21) and dense sampling (0.23) with single-action Support Vector Regression (SVR) employing a linear kernel. The proposed feature representation slightly improved the result of DCT and achieved the best value of 0.41. The promotion was limited, likely due to the simple segmentation of videos according to the equal length strategy, and the synchronization of segments between different action instances that were not considered. It is noted that the performance of the original SSM feature method was clearly decreased for this dataset, and the best result obtained was 0.19, which was inferior to the low-level feature of STIP and dense sampling. This indicated that the original SSM feature is unsuitable for fine-grained feature representation of long-duration video sequences, likely because it accumulated all joints' relative position offsets between each two successive frames of a video sequence. However, the motion data of each joint and joint relationship important for describing the intrinsic characteristic of different contained actions were completely ignored. We also compared the all-action evaluation and single-action evaluation methods, SVR regression, and the RR regression method using this dataset. The comparison of different evaluation methods is presented in Table 2. In terms of the best rank correlation coefficient obtained, the predicted score of each testing video for the MIT figure skating dataset is illustrated in Figure 8.

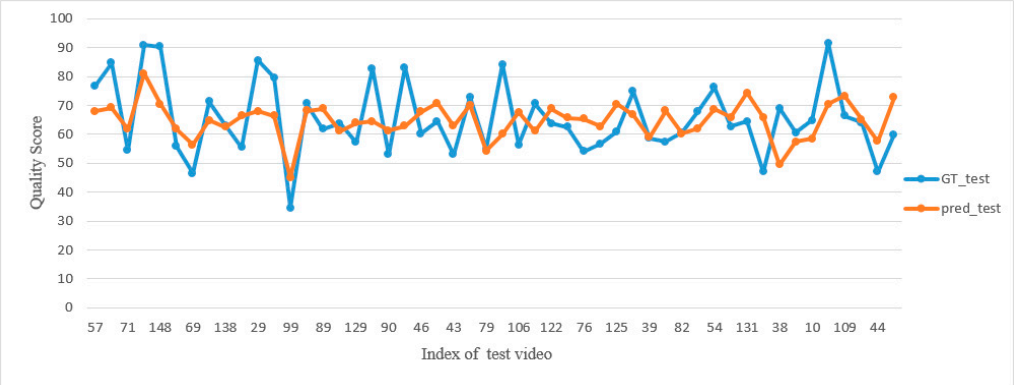



**Figure 8.** The predicted scores of test videos in terms of the best rank correlation coefficient obtained for the MIT figure skating dataset.

### *4.4. Results on the UNLV Vault Dataset*

The UNLV vault dataset comprises 176 videos of gymnastic vaulting captured from five international competitions of the Summer and Winter Olympics on YouTube. The average length is 75 frames per vault video. The frame size is 320 × 240. The ground-truth score of each vaulting video is freely obtained by extracting the judge's score that is publicly released in sports footage. It ranges from 12.30 to 16.87, and is determined by the sum of the "Execution" score and "Difficulty" score. Although a short average length and relatively simple actions are contained in vault videos compared with those of diving and figure skating, severe view changes exist in this dataset due to the different broadcast configurations employed for different events. Additionally, the low-resolution of some broadcasting videos make them more difficult to employ for action quality scoring.

We followed the evaluation protocol in Reference [27], which consists of a fixed split of the dataset, where 120 videos were chosen as training samples and the rest of the 56 videos were selected as testing samples. The comparison of our feature method with other reviewed state-of-the-art feature methods and different evaluation methods is presented in Table 3.

**Table 3.** Quality assessment results for the UNLV vault dataset. The rank correlation coefficient between the predicted results and the ground truth is presented.

| | STIP | Dense | Skeleton + DCT | Skeleton + SSM | Our Method |
|---|---|---|---|---|---|
| **All-action SVR-Linear** | 0.05 | 0.10 | 0.15 | 0.09 | **0.17** |
| **Single-action SVR-Linear** | 0.13 | 0.16 | 0.41 | 0.33 | **0.47** |
| **Single-action Ridge Regression** | 0.11 | 0.12 | 0.35 | 0.31 | **0.39** |

From Table 3, it can be concluded that skeleton-based feature representations exhibited a significantly greater strength than low-level features of STIP and dense sampling under most of the different evaluation strategies used for vault video assessment. In skeleton data-based feature representation, the best mean rank correlation coefficients of the DCT transformation feature, SSM feature, and our proposed feature were 0.41, 0.33, and 0.47, respectively. It outperformed STIP (0.13) and dense sampling (0.16) with single-action SVR using a linear kernel. The proposed feature representation improved the performance and achieved the best result of 0.47. In the comparison of different evaluation methods, as is shown in Table 3, it could also be found that, even though the performance difference between single-action evaluation and all-action evaluation under low-level STIP and dense sampling features is not clear, singe-action evaluation significantly improved the performance of skeleton-based feature representations. It is noted that the ridge regression obtained better results for this dataset than for the diving and figure skating datasets. The predicted score of each testing video for this dataset is illustrated in Figure 9.

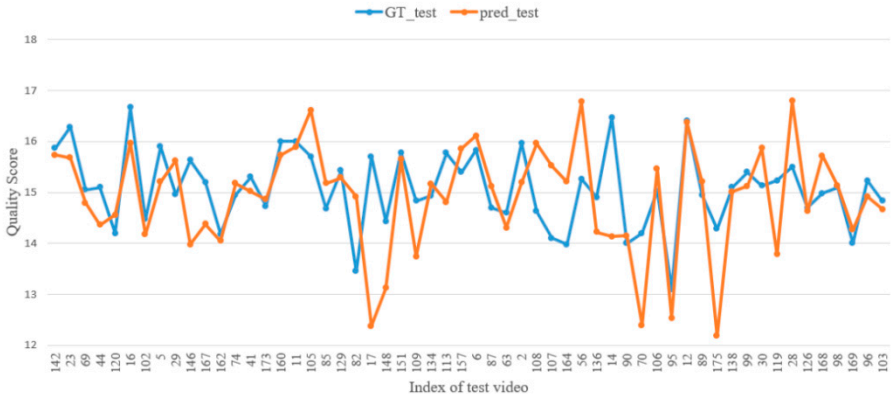



**Figure 9.** The predicted scores of test videos for the UNLV vault dataset.

### *4.5. Comparison with State-of-the-Art Feature Methods*

We also compared our method with other state-of-the-art feature methods for the MIT diving, MIT figure skating, and UNLV vault dataset, as shown in Table 4.

**Table 4.** Comparison results of the proposed method with the baseline and other relevant methods.

| | MIT Diving Dataset | MIT Figure Skating Dataset | UNLV Vaul Dataset |
|---|---|---|---|
| **STIP feature method [1]** | 0.10 | 0.21 | – |
| **Pose+DFT method [1]** | 0.27 | 0.31 | – |
| **Pose+DCT method [1]** | 0.41 | 0.35 | – |
| **ApEn method [15]** | 0.45 | – | – |
| **Our approach** | **0.52** | **0.41** | **0.47** |

As a baseline method, Pirsiavash et al. [1] proposed a general learning-based framework to assess the quality of human-based actions from videos. They proposed using DFT and DCT transformation of original coordinates of human body joints detected by the Flexible Parts Model [31], and selected k low frequency coefficients to represent human actions. They evaluated the proposed feature method using the MIT diving and figure skating datasets and compared the performance with two kinds of low-level features, namely the STIP feature and DFT pose feature. For the MIT diving dataset, the best results of the STIP feature were 0.07 with the support vector regression and 0.10 with ridge regression. The DCT pose feature achieved the best result of 0.41, which was superior to the STIP feature value of 0.10 and DFT pose feature value of 0.27. For the MIT figure skating dataset, the DCT transformation-based pose feature obtained the result of 0.35, which was better than the STIP feature value of 0.21 and DFT transformation pose feature value of 0.31. Venkataraman et al. [15] investigated approximate entropy-based (ApEn) feature representation for segmenting untrimmed motion capture data, and only evaluated their method for assessing the quality of diving actions included in the MIT Olympic Scoring dataset. They reported the best result of 0.45 for the diving dataset, which showed that a 10% improvement was achieved in the rank correlation coefficient when compared to Reference [1].

As shown in Table 4, from the results on MIT diving and figure skating of the above-mentioned handcrafted features, we can conclude that our proposed feature method achieved the best rank correlation coefficients of 0.52 and 0.41, respectively, for the two actions. It is believed that the proposed feature representation can better encode the dynamics of human motion for the fine-grained quality assessment of human actions than the reviewed state-of-the-art feature methods by capturing the self-similarity patterns from joint trajectories and joint displacement sequences.

## 5. Conclusions

In this paper, we have proposed an integrated category classification and regression-based evaluation framework for fine-grained human action quality assessment. In this framework, for action classification, the local motion patterns of body joint-based feature representation are extracted to train the discriminative classifier. The output of the classifier is used to supervise the quality assessment process in the testing stage. To deal with intra-class variations and acquire effective dynamic representation, the semantic pose feature captured from the self-similarity matrix of joint trajectories and joint displacement sequences is developed. A class-specific learning algorithm is employed to build an evaluation function for each action category. The experimental results show improvements for both the diving and figure skating datasets in comparison with other handcrafted feature methods.

The limitations of our proposed method include the fact that the segmentation of long videos has simply been considered and the synchronization of segments has not been researched. Our method is more suitable for assessing well-segmented action instances. When complex activities rather than actions are contained in videos with a long-time duration, the quality score is strongly affected. In future studies, semantic segmentation and alignment methods will be addressed to promote the practical application of the proposed framework.

**Author Contributions:** Conceptualization, Q.L. Methodology, Q.L. Software, Q.L. Validation, Q.L., H.-B.Z., and J.-X.D. Formal analysis, Q.L. Investigation, Q.L. Resources, Q.L. Data curation, T.-C.H. Writing—original draft preparation, Q.L. Writing—review and editing, Q.L. Visualization, T.-C.H. Supervision, J.-X.D. Project administration, C.-C.C. Funding acquisition, H.-B.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** 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), supported this work.

**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, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 556–571.
2. Lei, Q.; Du, J.-X.; Zhang, H.-B.; Ye, S.; Chen, D.-S. A Survey of Vision-Based Human Action Evaluation Methods. *Sensors* **2019**, *19*, 4129. [CrossRef] [PubMed]
3. 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.
4. 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.
5. 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.
6. 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.
7. 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]
8. 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]

9. 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]
10. Laraba, S.; Tilmanne, J. Dance performance evaluation using hidden markov models. *Comput. Animat. Virtual Worlds* **2016**, *27*, 321–329. [CrossRef]
11. Barnachon, M.; Boufama, B.; Guillou, E. A real-time system for motion retrieval and interpretation. *Pattern Recognit. Lett.* **2013**, *34*, 1789–1798. [CrossRef]
12. 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] [PubMed]
13. 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]
14. 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]
15. Venkataraman, V.; Vlachos, I.; Turaga, P. Dynamical Regularity for Action Analysis. In Proceedings of the 26th British Machine Vision Conference, Swansea, UK, 7–10 September 2015; British Machine Vision Association: Swansea, Wales, 2015; pp. 67.1–67.12.
16. Vicente, I.S.; Kyrki, V.; Kragic, D.; Larsson, M. Action recognition and understanding through motor primitives. *Adv. Robot.* **2007**, *21*, 1687–1707. [CrossRef]
17. 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]
18. 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]
19. 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]
20. 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]
21. 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.
22. Zanfir, M.; Leordeanu, M.; Sminchisescu, C. The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 2752–2759.
23. Nowozin, S.; Shotton, J. Action points: A representation for low-latency online human action recognition. *Mark. Health Serv.* **2013**, *32*, 3–5.
24. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In Proceedings of the International Conference on Pattern Recognition, Cambridge, UK, 23–26 August 2004; pp. 32–36.
25. Junejo, I.N.; Dexter, E.; Laptev, I. View-independent action recognition from temporal self-similarities. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 172–185. [CrossRef] [PubMed]
26. MIT Olympic Scoring Dataset. Available online: https://www.csee.umbc.edu/~{}hpirsiav/quality.html (accessed on 23 January 2020).
27. UNLV Olympic Scoring Dataset. Available online: http://rtis.oit.unlv.edu/datasets.html (accessed on 23 January 2020).
28. 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.
29. 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.

30. Wang, H.; Ullah, M.M.; Klaser, A.; Laptev, I.; Schmid, C. Evaluation of Local Spatio-temporal Features for Action Recognition. In Proceedings of the British Machine Vision Conference, London, UK, 7–10 September 2009; pp. 1–10.
31. Yang, Y.; Ramanan, D. Articulated pose estimation with flexible mixtures-of-parts. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1385–1392.