electronics

Article

An Efficient Motion Registration Method Based on Self-Coordination and Self-Referential Normalization

Yuhao Ren [1,2], Bochao Zhang [1,2], Jing Chen [1,2], Liquan Guo [1,2,*] and Jiping Wang [1,*]

[1] School of Biomedical Engineering (Suzhou), Division of Life Science and Medicine, University of Science and Technology of China, Hefei 230026, China
[2] Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou 215163, China
* Correspondence: guolq@sibet.ac.cn (L.G.); wangjp@sibet.ac.cn (J.W.)

Abstract: Action quality assessment (AQA) is an important problem in computer vision applications. During human AQA, differences in body size or changes in position relative to the sensor may cause unwanted effects. We propose a motion registration method based on self-coordination (SC) and self-referential normalization (SRN). By establishing a coordinate system on the human body and using a part of the human body as a normalized reference standard to process the raw data, the standardization and distinguishability of the raw data are improved. To demonstrate the effectiveness of our method, we conducted experiments on KTH datasets. The experimental results show that the method improved the classification accuracy of the KNN-DTW network for KTH-5 from 82.46% to 87.72% and for KTH-4 from 89.47% to 94.74%, and it improved the classification accuracy of the tsai-MiniRocket network for KTH-5 from 91.29% to 93.86% and for KTH-4 from 94.74% to 97.90%. The results show that our method can reduce the above effects and improve the action classification accuracy of the action classification network. This study provides a new method and idea for improving the accuracy of AQA-related algorithms.

Keywords: computer vision; AQA; motion registration; action feature fusion method; SC; SRN

Citation: Ren, Y.; Zhang, B.; Chen, J.; Guo, L.; Wang, J. An Efficient Motion Registration Method Based on Self-Coordination and Self-Referential Normalization. *Electronics* **2022**, *11*, 3051. https://doi.org/10.3390/electronics11193051

Academic Editor: Donghyeon Cho

Received: 29 August 2022
Accepted: 22 September 2022
Published: 24 September 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Action quality assessment (AQA) is an important issue that has emerged in recent years in various computer vision applications [1]. It is the process of quantifying the execution of an action or calculating a score that represents the quality of the action execution. The aim is to allow computers to automatically quantify the execution of human actions and, further, provide interpretable feedback to improve human actions [2]. Since the quality of human movements directly depends on their dynamics, how to obtain an effective motion representation is the fundamental problem of AQA [3]. Motion registration, as a key means to solve this problem, achieves the purpose of observing or differentiating movements by extracting the motion characteristics of human motion. The general steps are: extracting human motion-related features in various environmental contexts, then further extracting and normalizing the motion feature information, and finally performing human motion recognition estimation. Different motion registration methods may not use the same feature information, such as RGB information, skeleton information, infrared images, etc. However, when evaluating various motion registration methods, there is more concern about whether the movements are more standardized and distinguishable after the registration; in general, whether the corresponding features of each movement are more obvious and make the movements more distinguishable. A more intuitive evaluation method is to send the data to the neural network after processing to see if and how much the classification accuracy has improved.

Existing AQA-related work reduces errors and improves accuracy by changing the network structure or increasing the amount of data and the number of recognition features. With these innovative approaches, the corresponding algorithms can be validated as effective for improving AQA accuracy. However, these methodological changes may also increase the complexity of the algorithm and the network burden to some extent. This manifests itself in higher requirements for the hardware environment, making the algorithms less widely applicable. Despite the continuous development of technology, these errors in AQA still exist and it is crucial to reduce them. As a new method of motion registration, our method can reduce the above errors to some extent. The goal of improving the accuracy of motion recognition or AQA can be achieved by choosing a different direction from the traditional one. This can reduce the negative effects mentioned earlier. The motion features are first formed using the distance variation in the human skeleton joint points in relation to the origin extracted in one second. Then, a self-referential normalization method is used to reduce the motion recognition errors caused by the variation in human body spacing and dimensions. A self-coordination method is also used to reduce the errors caused by unnecessary displacements relative to the sensor during human motion. We chose to conduct classification tests on the KTH dataset using two relatively simple and shallow networks, KNN-DTW and tsai-MiniRocket, to verify the effectiveness of the method in improving the classification accuracy of the two classification networks.

The rest of the paper is organized as follows: the second part summarizes and introduces related work in China and abroad, the third part describes the specific content and methods of the experiment, the fourth part presents and analyzes the experimental results, and the fifth part summarizes the whole experiment.

2. Related Work

To improve the accuracy of motion registration and, thus, the accuracy of AQA-related algorithms, different scholars have chosen different directions.

Existing network improvements. S. Suraj Prakash and A. Samit proposed a human action recognition method based on a local maximal difference image (LMDI) interest-point detection technique, a random projection tree with overlapping segmentation and modified voting scores [4]. Tran et al. used deep 3D convolutional networks (3D ConvNets) to extract action features [5]. G. Batchuluun et al. proposed a combination of a convolutional neural network (CNN) and long short-term memory (LSTM) for human action recognition [6]. Wang et al. proposed a new parameter-free spatio-temporal pool (STP) for video action recognition [7]. Zhang et al. developed a software-based sensor calibration algorithm to design a pose-based adaptive signal segmentation algorithm [8]. Tang et al. combined the ResC3D network and convolutional LSTM (ConvLSTM) into a new, refined fusion model architecture with a dynamic selection mechanism [9]. F. Zhong et al. used DSPNet for human pose estimation [10]. X. Xu et al. proposed to use RSC-Net for pose estimation [11]. Hao et al. introduced the Hyper-GN Neural Network (Hyper-GNN) for action recognition [12]. Zheng et al. proposed GarmentNet and SynthesisNet for pose estimation [13]. Li et al. added an orthogonal soft-code layer (OSL) to an action classification network [14]. Farabi et al. used a 34-layer (2 + 1) D convolutional neural network for AQA [15]. All of these methods are excellent methods of the moment. They start with the optimization of the network structure, and all make their own innovative improvements on the existing network. The accuracy of the algorithm is improved by adding a new module or replacing a module in the network. However, these methods have relatively high hardware requirements. In other words, it may not be possible to obtain better results when the hardware is not up to the requirements.

Dataset expansion. K. Nishi et al. constructed a dataset containing 10,076 images using a method that generates annotated depth images of body parts in various body shapes and poses [16]. W. Ren et al. used a hybrid fuzzy logic and machine learning approach to classify human poses lying on a bed using a dataset containing 19,800 annotated depth images [17]. With the development of technology, a large amount of data can be stored in the form of images, which makes it possible to improve the accuracy of action recognition by increasing the size of the dataset. However, in real life, the amount of data that can be captured in any application scenario is not large. Even if it were possible, the algorithm running time, as well as the operational burden on the network, can increase substantially.

Feature addition. D. C. Luvizon et al. jointly estimated 2D and 3D human poses from single-shot color images and obtained human action from video sequences for classification [18]. Abdallah Benzine et al. proposed a single-shot method for multi-person 3D human pose estimation in complex images [19]. Gedamu Kumie et al. used a new two-branch viewpoint action generation method based on an auxiliary conditional GAN to achieve arbitrary viewpoint human action recognition [20]. W. Ding et al. proposed a multi-feature and rule-based human pose recognition algorithm [21]. H. Wang et al. proposed a skeleton edge motion network (SEMN) for action recognition [22]. J. Zhu et al. proposed a human-centric modeling video action recognition framework for action recognition [23]. Chang et al. proposed a long-term video action recognition (LVAR) framework for continuous video action classification [24]. Angelini et al. used ActionX-Pose to extract low- and high-order features from body poses for pose recognition [25]. It is feasible to improve the accuracy of action estimation and recognition by adding new recognition features. However, like increasing the amount of data, introducing new features will also increase the computational burden of the network to a certain extent, which is relatively demanding on hardware. It may not be suitable for special environments (e.g., home, community, hospital, etc.).

Image retrieval. Today, large amounts of data are stored in image format. Content-based image retrieval from bulk databases has become an interesting research topic in the last decade. Most of the recent approaches use joint texture and color information. K. Nasim and S. Fekri-Ershad proposed a new approach based on weighted combination of color and texture features for image retrieval and obtained good performance [26]. In this field, our method can also play a role in eliminating errors and improving retrieval accuracy.

In summary, various experts and scholars share a common goal of improving the accuracy of algorithms related to action recognition. In traditional approaches, this can be achieved by extending the dataset, adding features, or improving existing network work. In contrast to the traditional direction of improving algorithm performance, our approach (SC + SRN) provides a new direction and idea for the improvement of the overall performance of action recognition- and classification-related algorithms, breaking the "ceiling" of existing action classification networks. In addition, as the volume of image data increases, the need for graphical retrieval of massive content-based data is also increasing. Our proposed method can also be used for image retrieval.

3. Approach

The main flow of this experiment is shown in Figure 1. First, we extracted 26 frames (fixed values) for each action video (frame rate of 25 fps) of the KTH dataset, enough frames to describe each action category, from 0–25 frames within 1 S from the beginning to the end. All extracted images were processed by existing opensource methods to obtain the corresponding human skeletal feature points in each image (23 feature points in total). Here, we used the opensource project "openpose" to extract the skeleton, a method that determines the skeletal feature points through a combination of heat and confidence mapping. Fourteen human skeleton feature points (top of head, neck, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, right ankle; see Figure 2 for details, except nine facial feature points, which are irrelevant for pose recognition) were selected in each frame. Then, the sample data were processed in three different ways: the raw unprocessed sample (Raw), which uses the default coordinate origin; SC, which fixes the origin of the feature point coordinate system to the human body (the new coordinate origin is shown in Figure 2); and SC

+ SRN, which uses self-coordination combined with our self-reference normalization. After that, we calculated the Euclidean distances of 14 feature points relative to the origin for each frame and arranged them in the temporal order of 26 frames for feature fusion. After processing and feature fusion, different actions were given different labels (1–6, indicating the correct classification of each action). Then, we used the KNN-DTW network and tsai-MiniRocket network to perform action classification tests. Finally, the obtained experimental results were unified, compared, analyzed, and summarized.

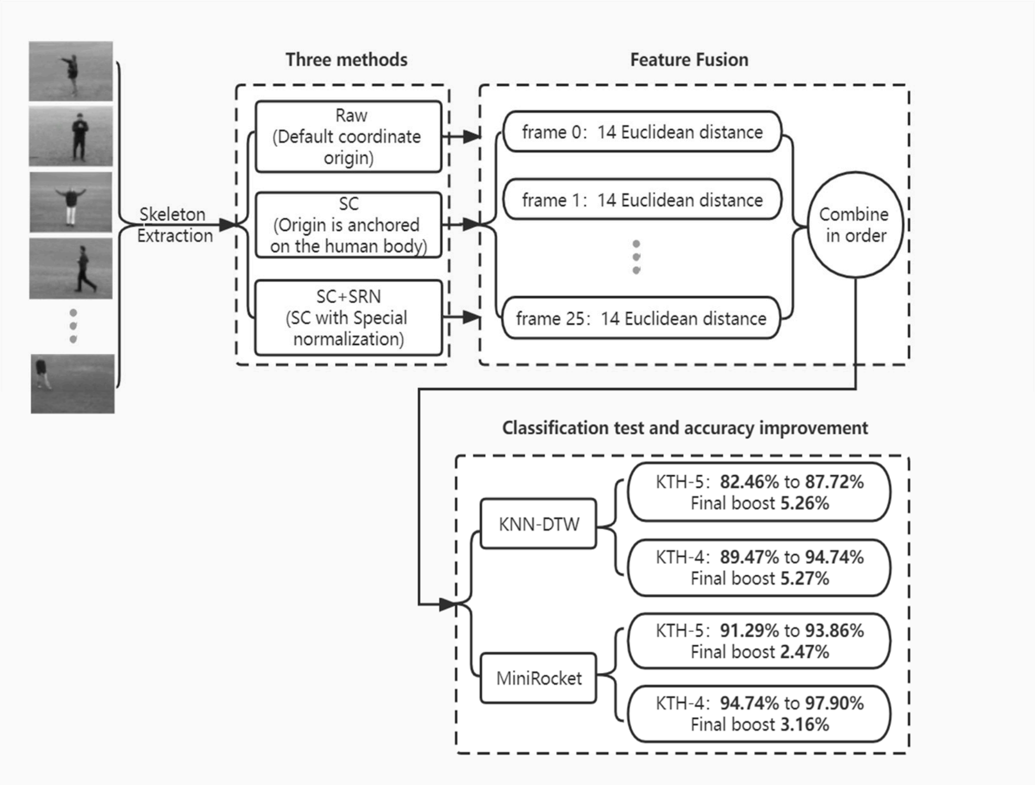

Three methods
Feature Fusion
Skeleton Extraction
Raw (Default coordinate origin)
SC (Origin is anchored on the human body)
SC+SRN (SC with Special normalization)
frame 0: 14 Euclidean distance
frame 1: 14 Euclidean distance
frame 25: 14 Euclidean distance
Combine in order
Classification test and accuracy improvement
KNN-DTW
MiniRocket
KTH-5: 82.46% to 87.72% Final boost 5.26%
KTH-4: 89.47% to 94.74% Final boost 5.27%
KTH-5: 91.29% to 93.86% Final boost 2.47%
KTH-4: 94.74% to 97.90% Final boost 3.16%

Figure 1. Flow chart for the experimental method. The proposed approach consists of three main components: handling, feature fusion, and classification testing. In addition, the accuracy improvement in the method for the classification tests is also listed.

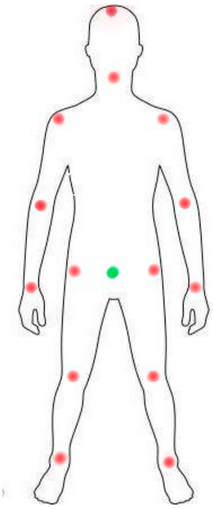


Figure 2. Schematic diagram of the location of the selected feature points (red dots are the 14 selected feature points; green dots are the anchor points of the SC).

3.1. SC

The core of the SC involves fixing the origin of the coordinate system on the human body. The purpose is to reduce the negative effects of pose recognition when the human body is unnecessarily displaced in the plane with respect to the sensor. This effect is expressed in the change in the displacement of each feature point relative to the origin of the picture coordinates. The central idea of the SC is to fix the origin of the coordinates in the body itself to reduce the above effects. The second of the four scenes in the KTH dataset simulates the above effect through the near and far transformation of the camera, which is the main reason for choosing the KTH dataset. We chose the human body center-of-gravity point as the anchor point (Figure 2, green dot). As the human body will keep the center of gravity stable during movement, it is reasonable to use this point as the anchor point. Although there is no center of gravity point among the extracted feature points, it can be calculated indirectly through other points with the following calculation formula:

$$X_{(x,y)} = \frac{\left(X_{L(x,y)} + X_{R(x,y)}\right)}{2} \tag{1}$$

where $X_{(x,y)}$ is the coordinate of the anchor point, $X_{L(x,y)}$ is the coordinate of the left hip, and $X_{R(x,y)}$ is the coordinate of the right hip.

3.2. Normalization

The purpose of normalizing the motion data is to reduce the effect of the difference in body size or the front-to-back displacement relative to the sensor during motion in pose recognition. The common methods of normalization include standard normalization and maximum–minimum normalization. Descriptions of the two common normalization methods and our proposed SRN method are provided below.

3.2.1. Standard Normalization

The raw data are normalized to have a mean of 0 and a variance of 1. The normalization formula is as follows:

$$X^* = \frac{X - \mu}{\delta} \tag{2}$$

where μ is the mean of all sample data, δ is the standard deviation of all sample data, X is the raw data, and X^* is the normalized data.

3.2.2. Maximum–Minimum Normalization

The raw data are converted to the range [0,1] according to the linearization method with the following normalization formula:

$$X^* = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{3}$$

where $X_{\max}$ is the maximum value of the sample data, $X_{\min}$ is the minimum value of the sample data, X is the raw data, and X^* is the normalized data.

At the beginning of the experiment, we normalized the data with maximum–minimum normalization. However, the results were not satisfactory: the accuracy rate did not improve but decreased a lot. After summarizing and analyzing, we developed the following SRN method, which is very effective.

3.2.3. SRN

SRN is a normalization method that uses itself or a part of itself as a reference. Considering the principle of normalization and our experimental needs, we finally chose 10 times the distance between the top of head and the neck (approximately equal to the length of the human head) as the reference. The reason is that the head is always rigid during human motion and the distance between the above two points is relatively stable in the KTH dataset. This distance was multiplied by a factor of 10 to normalize each data point in the range of (0,1). This is so that, when the origin of the coordinate system is fixed on the human body, the distance from any feature point to the origin will not exceed 10 times the distance between the top of the head and the neck. The formula for this normalization method is as follows:

$$X^* = \frac{X}{10\left|X_h - X_n\right|} \tag{4}$$

where X_h is the distance from the key point of the head to the origin, X_n is the distance from the key point of the neck to the origin, X is the raw data, and X^* is the normalized data.

3.3. Skeleton Extraction

″openpose″ is an open source and mature method for human skeleton feature extraction. It obtains the joint heat map and confidence map by analyzing the human body in the image. The maximum possible position of the joint is then calculated as the position of the human joint. In this way, the human skeleton can be obtained.

In this study, we used the ″light-openpose″ method to extract the human skeleton, which was previously developed by members of our lab. This method is different from ″openpose″. It is more lightweight and, therefore, can reduce the hardware requirements and increase the running speed to accommodate lightweight system development needs. However, this sacrifices a certain degree of accuracy in the skeleton joint point recognition. This may have been a factor that affected the results of our experiments and is a direction that it may be possible to improve in the future. Since we were targeting an application range for homes and communities, it was reasonable to choose “light-openpose”. By loading the video clips into the model, we were able to obtain data on 23 key points of the human skeleton.

3.4. Feature Fusion

The 14 feature points (red points) and the coordinate origin (green points) selected are shown in Figure 2. In the process of feature fusion, the Euclidean distance of 14 feature points in each frame relative to the origin of the coordinate system was first calculated. Then, they were arranged in the order of head top, neck, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, right hip, left knee, right knee, left ankle, and right ankle. Finally, the calculated series of distances were combined in the temporal order of 0–25 frames to complete the feature fusion of an action (refer to the feature fusion section in Figure 1). After this feature fusion, the spatial information between feature points was retained and the temporal information for the consecutive actions was included to form a sample set. The average time that our method took for feature fusion was 12.08 ms (the average time spent per feature fusion action) on an AMD R7 CPU, which is less than the running time of a conventional network. This indicates that the additional time consumption required by our method is acceptable.

4. Experiments

4.1. Dataset and Networks

Dataset. We used the public dataset KTH for this experiment. The dataset contains a total of 2391 sets of data, including six actions (walking, boxing, handwaving, handclapping, jogging, running). Each action is undertaken by 25 characters in four different scenes, so there are 100 cases of each action. In total, as there 26 frame × 6 actions × 25 characters × 4 scenes, there are 15,600 pictures. When sampling the images in the video, the collected images with unsynchronized time and incomplete information between the various actions were filtered and discarded. Finally, five types of actions (KTH-5) were left after excluding running, in which the numbers of actions and samples were 82 cases for walking, 100 cases for boxing, 100 cases for handclapping, 99 cases for handwaving, and 75 cases for jogging. The training set and datasets were randomly composed in a 3:1 ratio. For KTH-5 classification, the training set consisted of 342 (62 + 75 + 74 + 75 + 56) samples and the test set consisted of 114 (20 + 25 + 25 + 25 + 19) samples. After analyzing the results of the KTH-5 classification, the jogging class was removed. Thus, for KTH-4 classification, the training set consisted of 286 (62 + 75 + 74 + 75) samples and the test set consisted of 95 (20 + 25 + 25 + 25) samples.

Networks. The KNN-DTW and tsai-MiniRocket networks used in the experiments are both from opensource collections shared by colleagues on the Internet. The difference is that we used our own feature fusion method and made corresponding changes to the network to suit the experimental needs. The reasons for choosing these two networks were as follows. DTW is the most prominent method used to describe similarities in time series data, and its combination with KNN could meet the experimental requirements. The MiniRocket network includes one of the most highly acclaimed opensource deep learning packages, named tsai, which met the experimental needs. The principle of the K-nearest neighbor (KNN) algorithm is to classify the test samples into the class with the largest number of all training samples in the range of the distance K from itself. Dynamic time winding (DTW) indicates the degree of similarity between two time series, and a smaller DTW distance indicates that the two series are more similar. The KNN-DTW network is a network that can classify sequences based on their similarity by replacing the radius K (the value of parameter K in the KNN) in the KNN with the DTW distance.

4.2. Results for KTH-5

First, the samples from KTH-5 (walking, jogging, boxing, waving, and clapping) were processed separately as follows: raw unprocessed samples (Raw), self-coordination (SC), and self- coordination + self-referential normalization (SC + SRN). Then, after feature fusion, the KNN-DTW with tsai-MiniRocket network was used for classification tests. The experimental results are as follows.

The results of the study on KNN-DTW networks using KTH-5 are shown in Tables 1–3. Our method significantly improved the prediction, recall, f1-score, and average classification accuracy for different kinds of actions. The proposed method had a significant positive effect on the entirety of the motion classification recognition.

Table 1. Classification results for Raw using KNN-DTW for KTH-5.

Action Category	Precision	Recall	f1-Score	Support
Walking	0.85	0.74	0.79	23
Boxing	0.96	0.96	0.96	25
Handclapping	0.84	0.81	0.82	26
Handwaving	0.80	0.80	0.80	25
Jogging	0.63	0.80	0.71	15

Accuracy	—	—	0.82	114

Table 2. Classification results for SC using KNN-DTW with KTH-5.

Action Category	Precision	Recall	f1-Score	Support
Walking	0.90	0.62	0.73	29
Boxing	0.92	1.00	0.96	23
Handclapping	0.96	0.83	0.89	29
Handwaving	0.88	0.96	0.92	23
Jogging	0.42	0.80	0.55	10
Accuracy	—	—	0.83	114

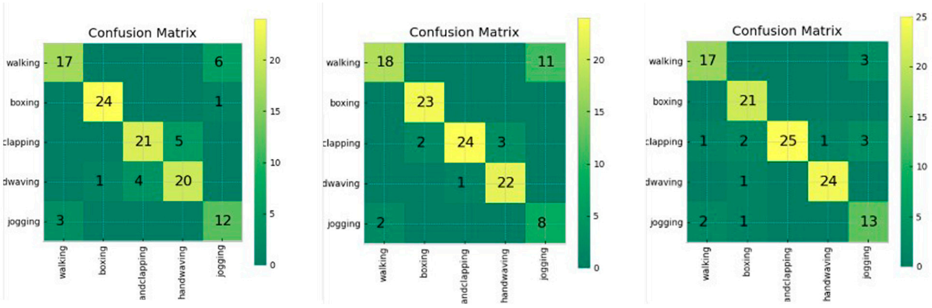
Table 3. Classification results for SC + SRN using KNN-DTW with KTH-5.

Action Category	Precision	Recall	f1-Score	Support
Walking	0.85	0.85	0.85	20
Boxing	0.84	1.00	0.91	21
Handclapping	1.00	0.78	0.88	32
Handwaving	0.96	0.96	0.96	25
Jogging	0.68	0.81	0.74	16
Accuracy	—	—	0.88	114

The overall classification accuracy scores for the two networks are shown in Table 4 and the covariance matrix corresponding to the KNN-DTW network test results is shown in Figure 3. For the KNN-DTW network, the action classification accuracy based on the raw data was 82.46%, and it reached 83.33% after adding the SC. After adding the SC + SRN, the classification accuracy was further improved to 87.72%, with a final improvement of 5.26%. For the tsai-MiniRocket network, the classification accuracy was 91.29% with the raw data and it reached 92.28% after adding the SC. After adding the SC + SRN, the classification accuracy was further improved to 93.86%, with a final improvement of 2.57%. To further examine the robustness of the method under different parameters, we also tested the KNN-DTW network with different radius K parameters. The results are shown in Figure 4. Our method significantly improved the network performance with all parameters.

Table 4. The classification accuracy for different methods and networks with KTH-5.

Method	KNN-DTW（%）	MiniRocket（%）
Raw	82.46	91.29
Our SC	83.33	92.98
Our SC + SRN	87.72	93.86



(a) (b) (c)

Figure 3. Covariance matrix for classification with KTH-5 using different methods of processing combined with KNN-DTW networks. (**a**) Raw, (**b**) SC, (**c**) SC + SRN.

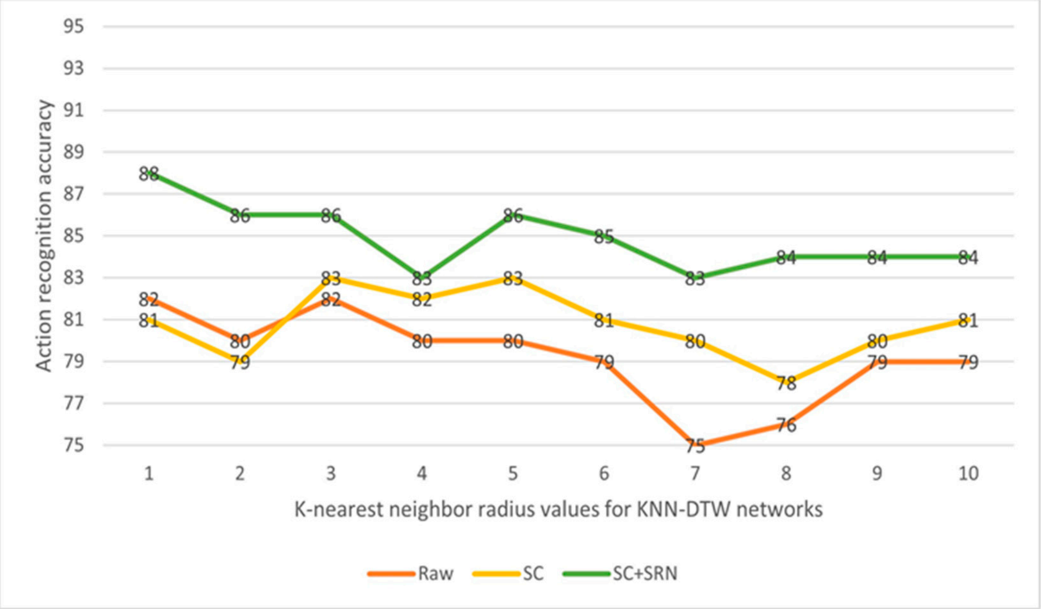


Figure 4. Classification results for KTH-5 with different radius K parameters using KNN-DTW network (the results are retained as integers).

4.3. Results for KTH-4

After observing the detailed results for KTH-5 classification using the KNN-DTW network. We found that the classification performance with the first four classes of actions (KTH-4) improved significantly after adding the SC, while the overall classification accuracy improved by only about 1%. This was due to the large decrease in the classification performance for the fifth category of actions. Referring to Figure 3a,b and Tables 1 and 2, this was because the network incorrectly predicted the fifth class of action (jogging) as the first class of action (walking). Combining the principles of KNN-DTW networks, we analyzed the reasons for this result. The use of SC reduced the magnitude of the change in the distance between each feature point and the origin, making the DTW distance between the two types of actions smaller with our feature fusion method. This reduced the final DTW distance variability between the two types of action sequences, resulting in the KNN-DTW network not being able to distinguish them significantly. This was caused by the misfit between the feature fusion method and the KNN-DTW network. Therefore, to verify the validity of the method and to reduce the effect of this misalignment, we removed the fifth category of actions. The remaining four categories of action samples (walking, punching, waving, and clapping) in the KTH dataset were tested for classification, and the results are as follows.

Tables 5–7 show the experimental results using the KNN-DTW network. The prediction, recall, f1 score, and average classification accuracy for different kinds of actions were consistent with the results for KTH-4. This indicates that our proposed method also had a significant positive effect on the recognition and classification performance of KTH-4 with KNN-DTW networks.

Table 5. Classification results for Raw using KNN-DTW with KTH-4.

Action Category	Precision	Recall	f1-Score	Support
Walking	1.00	1.00	1.00	20
Boxing	0.96	1.00	0.98	24
Handclapping	0.84	0.81	0.82	26
Handwaving	0.80	0.80	0.80	25
Accuracy	—	—	0.89	95

Table 6. Classification results for SC using KNN-DTW with KTH-4.

Action Category	Precision	Recall	f1-Score	Support
Walking	1.00	1.00	1.00	20
Boxing	0.88	1.00	0.94	22
Handclapping	0.96	0.83	0.89	29
Handwaving	0.88	0.92	0.90	24
Accuracy	—	—	0.93	95

Table 7. Classification results for SC + SRN using KNN-DTW with KTH-4.

Action Category	Precision	Recall	f1-Score	Support
Walking	1.00	1.00	1.00	20
Boxing	0.92	1.00	0.96	23
Handclapping	0.96	0.86	0.91	28
Handwaving	0.92	0.96	0.94	24
Accuracy	—	—	0.95	95

The overall classification accuracy scores for the two networks are shown in Table 8 and the covariance matrix corresponding to the KNN-DTW network test results are shown in Figure 5. With the KNN-DTW network, the classification accuracy based on the raw data was 89.47%. After adding the SC, the classification accuracy was significantly improved to 92.63%. After adding the SC + SRN, the classification accuracy was further improved to 94.74%, with an overall improvement of 5.27%. With the tsai-MiniRocket network, the classification accuracy for the raw data was 94.74%, It reached 96.84% after adding the SC and further increased to 97.90% after adding the SC + SRN, with an overall improvement of 3.16%. To further examine the robustness of the method under different parameters, we also tested the KNN-DTW network with different radius K parameters. The results are shown in Figure 6. This shows that our method could still significantly improve the network performance with all parameters. In addition, our proposed method maintained high and relatively stable accuracy even when the classification accuracy of the raw data decreased significantly at larger radius K.

Table 8. Classification accuracy for different methods and networks with KTH-4.

Method	KNN-DTW (%)	MiniRocket (%)
Raw	89.47	94.74
Our SC	92.63	96.84
Our SC + SRN	94.74	97.90

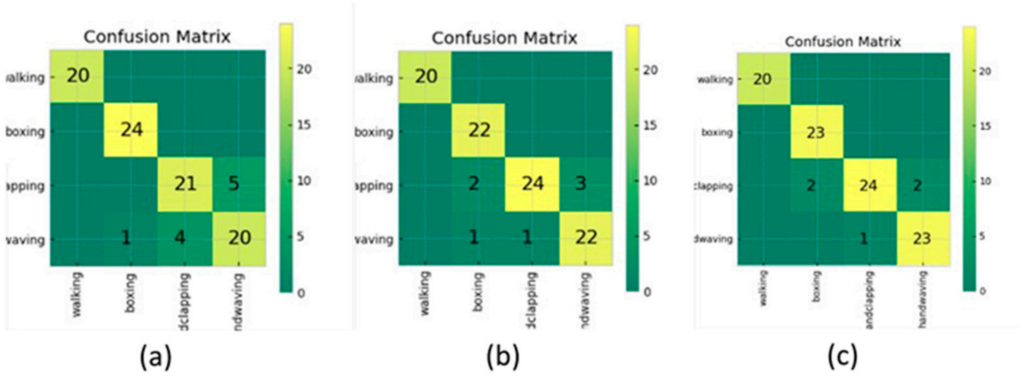


Figure 5. Covariance matrix for classification with KTH-4 using different methods of processing combined with KNN-DTW networks. (**a**) Raw, (**b**) SC, (**c**) SC + SRN.

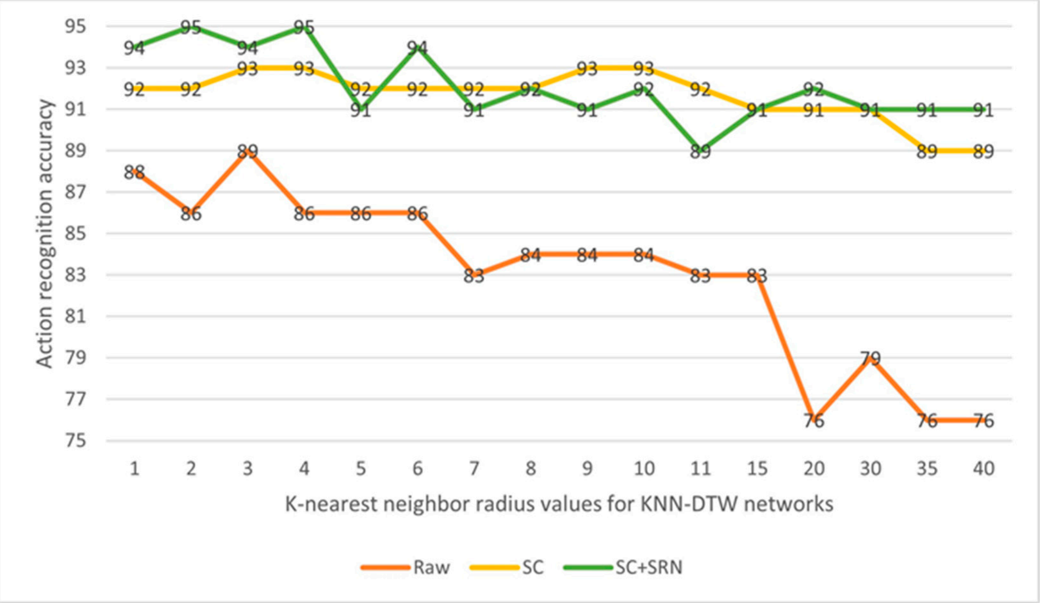


Figure 6. Classification results for KTH-4 with different radius k parameters using the KNN-DTW network (the results are retained as integers).

In summary, unlike approaches that extend datasets, add features, or improve existing networks, our motion registration method (SC + SRN) does not require large amounts of data or complex network structures. Simply combining it with a simple shallow network (e.g., the KNN-DTW network or tsai-MiniRocket network) can significantly improve the KTH-5 and KTH-4 action classification accuracy.

5. Conclusions

In this paper, a motion registration method based on self-coordination and self-reference normalization is proposed. The normality and distinguishability of different motion features are enhanced, and the effects caused by body size differences or changes in position relative to the sensor are reduced. Thus, the accuracy of motion registration and the accuracy of action classification recognition are improved. Using the KTH dataset, we validated the effectiveness of the method for improving the classification accuracy of action classification networks (KNN-DTW and tsai-MiniRocket). Our method cannot determine the classification accuracy of existing classification networks, but what we can improve is the original upper limit of classification networks. Differently from the traditional direction of improving algorithm performance, our method enhances the normality of raw data and provides a new direction and idea for improving the overall performance of action recognition- and classification-related algorithms. Additionally, the method also has some prospects for application in massive database image retrieval, at least in image retrieval related to people. Our method can reduce the variability in the same pose when different people or the same person is in different positions in the image. Thus, the accuracy of image retrieval can be improved and the relative search speed can be even further increased. In the future, we intend to continue to use this method in research on AQA-related algorithms to further test its applicability and feasibility and explore and broaden its scope of application.

Author Contributions: Writing—original draft preparation, Y.R.; validation, Y.R.; writing—review and editing, L.G., J.W., and J.C.; methodology, J.W.; formal analysis, B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by SUZHOU SCIENCE AND TECHNOLOGY PLAN, grant number SS201851 and Suzhou Special Technical Project for Diagnosis and Treatment of Key Clinical Diseases, grant number LCZX201931.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this paper are all available on request from the author.

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

References

1. 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.
2. Parmar, P.; Morris, B.T. Action quality assessment across multiple actions. *IEEE Winter Conf. Appl. Comput. Vis.* **2018**, 1468–1476. https://doi.org/10.1109/wacv.2019.0016.
3. Lei, Q.; Hongbo, Z.; Jixiang, D. Temporal Attention Learning for Action Quality Assessment in Sports Video. *Signal Image Video Process.* **2021**, *15*, 1575–1583.
4. Sahoo, S.P.; Ari, S. . On an Algorithm for Human Action Recognition. *Expert Syst. Appl.* **2019**, *115*, 524–534.
5. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. *IEEE Int. Conf. Comput. Vis.* **2015**, *2015*, 4489–4497.
6. Batchuluun, G.; Nguyen, D.T.; Pham, T.D.; Park, C.; Park, K.R. Action Recognition from Thermal Videos. *IEEE Access* **2019**, *7*, 103893–103917.
7. Wang, J.; Shao, Z.; Huang, X.; Lu, T.; Lv, X. Spatial–temporal Pooling for Action Recognition in Videos. *Neurocomputing* **2021**, *451*, 265–278.
8. Zhang, S.; Callaghan, V. Real-time Human Posture Recognition Using an Adaptive Hybrid Classifier. *Int. J. Mach. Learn. Cybern.* **2020**, *12*, 489–499.
9. Tang, X.; Yan, Z.; Peng, J.; Hao, B.; Wang, H.; Li, J. Selective Spatiotemporal Features Learning for Dynamic Gesture Recognition. *Expert Syst. Appl.* **2021**, *169*, 114499.
10. Zhong, F.; Li, M.; Zhang, K.; Hu, J.; Liu, L. DSPNet: A Low Computational-cost Network for Human Pose Estimation. *Neurocomputing* **2021**, *423*, 327–335.
11. Xu, X.; Chen, H.; Moreno-Noguer, F.; Jeni, L.A.; De la Torre, F. 3D Human Pose, Shape and Texture from Low-Resolution Images and Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4490–4504.
12. Hao, X.; Li, J.; Guo, Y.; Jiang, T.; Yu, M. Hypergraph Neural Network for Skeleton-Based Action Recognition. *IEEE Trans. Image Process.***2021**, *30*, 2263–2275.
13. Zheng, H.; Chen, L.; Xu, C.; Luo, J. Pose Flow Learning from Person Images for Pose Guided Synthesis. *IEEE Trans. Image Process.* **2021**, *30*, 1898–1909.
14. Li, X.; Chang, D.; Ma, Z.; Tan, Z.-H.; Xue, J.-H.; Cao, J.; Yu, J.; Guo, J. OSLNet: Deep Small-Sample Classification with an Orthogonal Softmax Layer. *IEEE Trans. Image Process.* **2020**, *20*, 6482–6495.
15. Farabi, S.; Himel, H.H.; Gazzali, F.; Hasan, B.; Kabir, M.; Farazi, M. Improving Action Quality Assessment using ResNets and Weighted Aggregation. *arXiv.* **2021.** Available online: https://doi.org/10.48550/arXiv.2102.10555. (accessed on 21th February, 2021).
16. Nishi, K.; Miura, J. Generation of human depth images with body part labels for complex human pose recognition. *Pattern Recognit.* **2017**, *71*, 402–413.
17. Ren, W.; Ma, O.; Ji, H.; Liu, X. Human Posture Recognition Using a Hybrid of Fuzzy Logic and Machine Learning Approaches. *IEEE Access* **2020**, *8*, 135628–135639.
18. Luvizon, D.C.; Picard, D.; Tabia, H. Multi-Task Deep Learning for Real-Time 3D Human Pose Estimation and Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 2752–2764.
19. Benzine, A.; Luvison, B.; Pham, Q.C. Achard, Single-shot 3D multi-person pose estimation in complex images. *Pattern Recognit.* **2021**, *112*, 107534.
20. Gedamu, K.; Ji, Y.; Yang, Y.; Gao, L.; Shen, H.T. Arbitrary-view human action recognition via novel-view action generation. *Pattern Recognit.* **2021**, *118*, 108043.
21. Ding, W.; Hu, B.; Liu, H.; Wang, X.; Huang, X. Human posture recognition based on multiple features and rule learning. *Int. J. Mach. Learn. Cybern.* **2020**, *11*, 2529–2540.
22. Wang, H.; Yu, B.; Xia, K.; Li, J.; Zuo, X. Skeleton Edge Motion Networks for Human Action Recognition. *Neurocomputing* **2021**, *423*, 1–12.
23. Zhu, J.; Zou, W.; Zhu, Z.; Xu, L. Action Machine: Toward Person-Centric Action Recognition in Videos. *IEEE Signal Process. Lett.* **2019**, *26*, 1633–1637.
24. Chang, Y.L.; Chan, C.S.; Remagnino, P. Action Recognition on Continuous Video. *Neural Comput. Appl.* **2020**, *33*, 1233–1243.
25. Angelini, F.; Fu, Z.; Long, Y.; Shao, L.; Naqvi, S.M. 2D Pose-Based Real-Time Human Action Recognition with Occlusion-Handling. *IEEE Trans. Multimed.* **2020**, *22*, 1433–1446.

26. Kayhan, N.; Fekri-Ershad, S. Content based image retrieval based on weighted fusion of texture and color features derived from modified local binary patterns and local neighborhood difference patterns. *Multimed. Tools Appl.* **2021**, *80*, 32763–32790.