symmetry

MDPI

Article

Human Action Recognition Using Bone Pair Descriptor and Distance Descriptor

Dawid Warchoł * and Tomasz Kapuściński

Department of Computer and Control Engineering, Faculty of Electrical and Computer Engineering, Rzeszów University of Technology, W. Pola 2, 35-959 Rzeszów, Poland; tomekkap@kia.prz.edu.pl

* Correspondence: dawwar@kia.prz.edu.pl; Tel.: +48-17-865-1614

Received: 6 September 2020; Accepted: 21 September 2020; Published: 23 September 2020



Abstract: The paper presents a method for the recognition of human actions based on skeletal data. A novel Bone Pair Descriptor is proposed, which encodes the angular relations between pairs of bones. Its features are combined with Distance Descriptor, previously used for hand posture recognition, which describes relationships between distances of skeletal joints. Five different time series classification methods are tested. The selection of features, input joints, and bones is performed. The experiments are conducted using person-independent validation tests and a challenging, publicly available dataset of human actions. The proposed method is compared with other approaches found in the literature achieving relatively good results.

Keywords: human action recognition; skeletal data; Kinect; descriptors

1. Introduction

Automatic human action recognition is an important research topic in machine vision and learning. It allows us to understand the intentions of a human, which can be a useful information in video surveillance, detection of aggressive behavior, human–computer, and human–robot interaction. The development of low-cost devices such as the Microsoft Kinect sensor caused increased interest in recognition methods based on depth data, such as point clouds, depth maps, and skeletons [1–6]. The skeletal data consists of 3D coordinates of characteristic points of human body. The main advantage of this type of data, compared to color/gray images, point clouds, and depth maps, is its size. A typical human action recorded as a sequence of 50 skeletons has the size of 4 KB, whereas a similar image or depth map sequence has the size up to several dozen MB. It makes a huge difference in the case of large datasets used for classifier training. Although the significant progress has been made in human action recognition, the existing algorithms are still far from being perfect, especially if a person performing actions is not present in the training dataset.

In this paper, a problem of human action recognition based on skeletal data is tackled. Our approach is based on Distance Descriptor and novel Bone Pair Descriptor which is a modification of a method previously developed for static hand posture recognition. The experimental tests are performed using five classifiers and different configurations of features. The main contributions of this paper are as follows.

1. The development of the Bone Pair Descriptor.
2. The application of the Distance Descriptor to the human action recognition problem.
3. The original experiments with the selection of joints, bones, and features.

The paper is organized as follows. The related works are discussed in Section 2. The proposed descriptors used for human action recognition are presented in Section 3. The used dataset, classifiers,

hardware, and performed experiments are characterized in Section 4. Section 5 contains the conclusions and the plans for the future work related to this subject.

2. Related Work

In recent years, several reviews on vision-based human action recognition have been published. Their authors attempt to classify different techniques.

Most often, the methods are divided into solutions based on handcrafted features or deep learning approaches. Features are calculated from color images, depth maps, skeletons, or by combining many modalities.

Many algorithms use 3D data most often in the form of depth maps. They facilitate the segmentation of the human silhouette in the case of a complex and heterogeneous background. In some approaches, spatial data is projected onto three orthogonal planes, corresponding to the front, side, and top views [7–12]. Depth motion maps (DMM), in various variants, are used in [8,10,11]. There are also solutions in which descriptors are built based on normals to the surface spanned on a three-dimensional point cloud [13,14]. In turn, the works in [15,16] describe algorithms based on the detection of points or regions of interest in a depth image.

Solutions using skeletal data can be classified into trajectory and pose-based [17]. The first group includes works, in which multivariate time series, obtained from space-time joints trajectories, are recognized [18,19]. In the second group, features describing the relationship between the skeleton elements, that determine a specific pose, are employed. The works in [1,2,20,21] use joints locations, angles between them, or more complex relationships linking body parts.

The solutions proposed in [22–24] combine the skeletal data with local features extracted from depth images in the neighborhood of the projected joints. In [25], skeletal data were used together with histogram of oriented gradients (HOG) descriptors calculated for regions of interest defined in RGB and depth images.

According to the authors of [26], deep learning solutions use spatio-temporal images + convolutional neural networks (CNN), recurrent neural networks (RNN), multiple-stream networks (MSN), and hybrid solutions (CNN + RNN).

Many methods exist for coding space-time data into two-dimensional images, which are then processed by convolutional networks. Several versions of chronological spatial-temporal images are proposed. In these solutions, the columns represent the coded spatial configuration of the nodes [27–29] or the derived features [5,30–32], and the rows correspond to the time domain. Other coding methods involve DMM calculated for three orthogonal planes [7] or motion history images derived from RGB data [33] or skeletal data [4,34].

In [35–38], RNN and long short-term memory network (LSTM) in various variants are used. The extension of the CNN network in the time domain is proposed in [39,40]. The paper [41] presents a hierarchical approach consisting of the decomposition of complex actions on simple ones. Hybrid solutions, with different variants of the CNN and LSTM networks connected sequentially, are used in [42–46]. In [47], the long-term recurrent convolutional network (LRCN) with jointly trained convolutional (spatial) and recursive (temporal) parts, is used.

Depending on the complexity, human activities can be classified into gestures, actions, interactions, and group collaborations. Therefore, in [48], methods were divided into single-layered and hierarchical. The first ones, designed to recognize simpler actions, use image sequences directly. In hierarchical approaches, complex activities are identified as combinations of simpler ones called subevents.

The problem is open and challenging. We propose the pose-based approach to human action recognition using only skeletal data and descriptors combining angular and positional relations of joints and bones corresponding to them.

3. Proposed Method

Our approach to classification of human actions is based on two descriptors: Distance Descriptor and Bone Pair Descriptor, which are characterized in the following subsections.

3.1. Distance Descriptor

The Distance Descriptor (DD) was first introduced in [49] for recognition of static hand postures. It encodes relationships between distances of skeletal joints from each other. Calculation of DD requires only 3D coordinates of joints. It does not use vectors or any other information. The descriptor can be computed for N joints by the following algorithm.

1. For each joint P_i, $1 \leq i \leq N$ do:
   (a) Calculate distances (Euclidean or city block) between the other joints P_j, $j \neq i$.
   (b) Sort joints P_j by the calculated distances from the closest to the farthest.
   (c) Assign consecutive integers a_{ij} to the sorted joints P_j, starting from 1.
2. Assemble a feature vector consisting of integer values assigned to the joints P_j in step 1c) in the following order: $[a_{12}, a_{13}, a_{14}, a_{15}, a_{21}, ..., a_{NN-1}]$.
3. Reduce the feature vector assembled in step 2 by adding together integers a corresponding to the same pair of indices i, j: $[a_{12} + a_{21}, a_{13} + a_{31}, ..., a_{N-1N} + a_{NN-1}]$.

Step 3 is performed not only to reduce the number of features. After this step, for each joint P_i, the algorithm determines not only which of the remaining joints P_j are its closest neighbors but also which of the joints P_j consider P_i as their closest neighbor. Finally, in order to normalize feature values to the interval [0–1], each of them is divided by $2(N-1)$.

The calculation of DD for the whole skeleton is time-consuming and not very effective in terms of classification accuracy. Therefore, an input set of joints should be selected from the whole skeleton.

3.2. Bone Pair Descriptor

The Bone Pair Descriptor (BPD) encodes the angular relations between particular pairs of bones. It is based on the Point Pair Descriptor (PPD) first introduced in [50] for recognition of static hand postures. PPD uses 3D joint coordinates, vectors pointed by fingers, and surface normals. BPD is a modification of PPD that uses bones as vectors allowing to describe human actions based only on skeletal data without surface normals. BPD can be calculated as follows. Let P_c be the skeleton central joint, b_c central vector associated with the joint P_c, P_i the i-th non-central joint, and b_i the vector associated with that joint (Figure 1). Vectors b_c and b_i coincide with a bone or a part of spine.

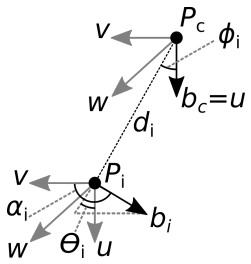

Figure 1. Construction of bone pair descriptor.

The relative position of vectors b_c and b_i is described by values α, ϕ, and Θ according to the Formulas (1)–(3) [51]:

$$\alpha_i = acos(v_i \cdot b_i) \tag{1}$$

$$\phi_i = acos\left(u \cdot \frac{d_i}{|d_i|}\right) \tag{2}$$

$$\Theta_i = atan\left(\frac{w_i \cdot b_i}{u \cdot b_i}\right) \tag{3}$$

where the vectors u, v_i, and w_i define the Darboux frame [52]:

$$u = b_c \tag{4}$$

$$v_i = \frac{d_i}{|d_i|} \times u \tag{5}$$

$$w_i = u \times v_i \tag{6}$$

with $\cdot$ denoting the scalar product and $\times$ denoting the vector product. Let N be the number od non-central joints. The Bone Pair Descriptor consists of $3N$ features calculated for each non-central joint using the Formulas (1)–(3):

$$V = [\alpha_1, \phi_1, \Theta_1, \alpha_2, \phi_2, \Theta_2, ..., \alpha_N, \phi_N, \Theta_N] \tag{7}$$

Finally, the features are normalized to the interval [0–1]. For this purpose, each feature is divided by its maximum possible value: π for features α, ϕ and 2π for feature Θ.

BPD requires the selection of central joint P_c, non-central joints P_i and joints determining vectors (bones) b_c and b_i from the whole skeleton.

The Matlab scripts for Distance Descriptor and Bone Pair Descriptor can be downloaded from our website [53].

4. Experiments

4.1. Dataset, Classifiers and Hardware

The experiments were performed using UTD-MHAD dataset [3] recorded by a Microsoft Kinect sensor. It contains 27 actions performed by 8 subjects (4 women and 4 men). Each subjects repeated each action 3 or 4 times, which is 861 sequences in total. The average action length is 68 frames. UTD-MHAD is a challenging dataset because of the large number of action classes and similarities between some of them.

The actions of UTD-MHAD dataset are listed below.

1. right arm swipe to the left
2. right arm swipe to the right
3. right hand wave
4. two hand front clap
5. right arm throw
6. cross arms in the chest
7. basketball shoot
8. right hand draw x
9. right hand draw circle (clockwise)
10. right hand draw circle (counter-clockwise)
11. draw triangle
12. bowling (right hand)
13. front boxing
14. baseball swing from right
15. tennis right hand forehand swing
16. arm curl (two arms)
17. tennis serve
18. two hand push
19. right hand knock on door
20. right hand catch an object

21. right hand pick up and throw
22. jogging in place
23. walking in place
24. sit to stand
25. stand to sit
26. forward lunge (left foot forward)
27. squat (two arms stretch out)

We used only skeletal data for action recognition. The whole skeleton of UTD-MHAD dataset consists of 20 joints as presented in Figure 2 (left image).

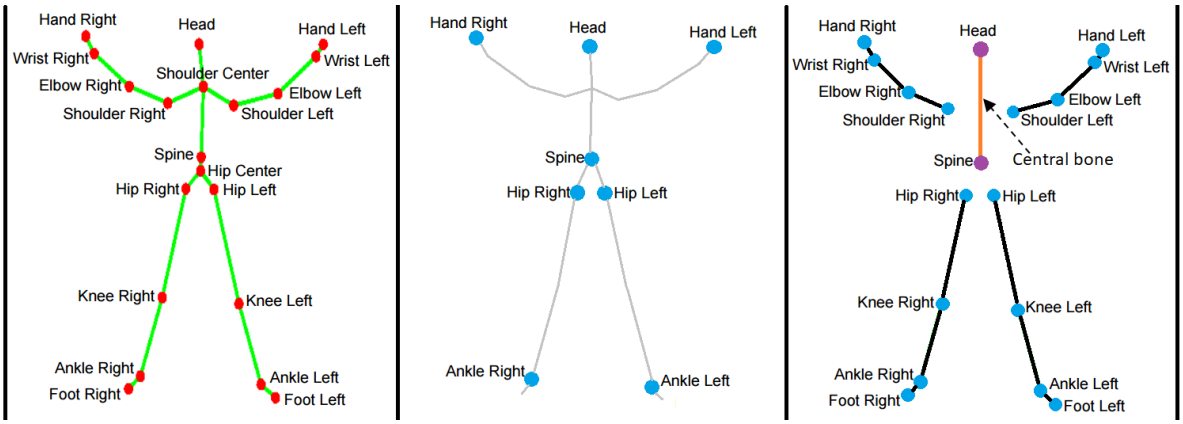

Figure 2. The whole skeleton of UTD-MHAD dataset (**left image**) [54], the subset of joints selected as an input for Distance Descriptor (DD) (**central image**), and the subset of joints selected as an input for Bone Pair Descriptor (BPD) (**right image**).

We selected the following subset of joints as an input for Distance Descriptor.

1. Hand Left
2. Hand Right
3. Shoulder Left
4. Shoulder Right
5. Head
6. Spine
7. Hip Left
8. Hip Right
9. Ankle Left
10. Ankle Right

This subset is shown in central image of Figure 2. The selected joints are presented as dots. We selected the following subset of bones as an input for Bone Pair Descriptor.

1. Spine—Head (central joints)
2. Elbow Right—Wrist Right
3. Wrist Right—Hand Right
4. Shoulder Right—Elbow Right
5. Elbow Left—Wrist Left
6. Wrist Left—Hand Left
7. Shoulder Left—Elbow Left
8. Hip Right—Knee Right
9. Knee Right—Ankle Right
10. Ankle Right—Foot Right
11. Hip Left—Knee Left

12. Knee Left—Ankle Left
13. Ankle Left—Foot Left

The selected non-central joints are presented as dots in the right image of Figure 2. The bones (vectors) are presented as lines connecting the joints. The dots at the ends of central bone are the central joints.

We selected those subsets of joints experimentally. We also tested different configurations; however, the presented subsets yielded the best results in terms of recognition rate and computation time.

For the evaluation of our method we used the protocol suggested by the authors of UTD-MHAD dataset [3]. Subjects 1, 3, 5, and 7 were used as a training set and subjects 2, 4, 6, and 8 were used as a testing set.

In our experiments we used five different classifiers: (1) dynamic time warping with Euclidean distance (DTW-euc), (2) dynamic time warping with city block distance (DTW-cb), (3) fully convolutional network (FCN), (4) bidirectional long short-term memory network (BiLSTM), and (5) LogDet divergence-based metric learning with triplet constraints (LDMLT). DTW-euc and DTW-cb are the classic methods for time series classification using dynamic programming to calculate distance of two nonlinearly aligned sequences. BiLSTM [55] and FCN [56] are the deep learning methods that are well known and have been successfully used, e.g., in speech recognition [57] and networking [58]. LDMLT is a relatively new method based on Mahalanobis distance learned using the so-called triplets constraints [59]. The output of DTW-euc, DTW-cb, and LDMLT is not a class label but a distance between two given sequences (each testing sequence have to be compared to each training sequence). Therefore, there is a need to apply K-nearest neighbors classifier searching for a class represented by the majority of K nearest neighbors. In our experiments we set K to 1.

The experiments were performed using Matlab R2019a software on a PC with Intel Core i7-4710HQ, 2.5 GHz CPU, and 16 GB RAM.

4.2. Experimental Results

We started our experiments with the comparison of the classifiers. In Table 1, we present the configuration of parameters and recognition rates for each classifier. The parameters were chosen experimentally, i.e., by changing their values and observing whether the recognition rate improved.

Table 1. Recognition rates [%] for various classifiers and their parameters.

Classifier	Parameter Name	Parameter Value	Recognition Rate [%]
BiLSTM	number of hidden units	150	76.7
	number of epochs	500	
	mini-batch size	45	
	initial learn rate	0.001	
DTW-cb	window size	5	81.4
FCN	number of layers	3	84.3
	number of epochs	2000	
	batch size	16	
DTW-euc	window size	5	86.1
LDMLT	triplets factor	20	92.1
	maximum cycle	15	
	alpha factor	5	

The LDMLT classifier yielded the best result with a large advantage over the other methods. Therefore, we used this classifier in the latter experiments.

We evaluated our method using feature vectors consisting of single DD, single BPD, and the concatenation of DD and BPD features. The results for various neighbors number K are shown

in Table 2. The highest recognition rate of single DD is 86.3% while single BPD achieved 87.2%. Combination of DD and BPD features led to more than 5 percentage points improvement yielding 92.6% accuracy for $K = 2$. This result confirms that the positional information of DD and the angular information of BPD complement each other improving overall recognition rate.

As the Bone Pair Descriptor consists of three different features (calculated for each non-central bone and the central bone), we tried removing one or two of them from the feature vectors. Θ turned out to be the least effective and its removal does not affect the highest recognition rate, which is still 92.6% for $K = 2$. Using two BPD features instead of three results in less time-consuming feature extraction, learning, and classification, and therefore DD + BPD (α, ϕ) can be considered the best configuration. We also tried calculating DD with city block distance instead of Euclidean, however the results were very similar (slightly lower).

Table 2. Recognition rates [%] for descriptors DD and BPD, various feature sets, and classifier neighbors number K. The classification was performed using LDMLT. The best results are highlighted.

	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
DD	86.3	83	82.8	79.1	79.3
BPD (all features)	87.2	86.5	84.9	83.5	81.6
DD + BPD (all features)	92.1	92.6	89.5	88.8	87.9
DD + BPD (ϕ, Θ)	91.9	91.6	89.8	88.4	86.5
DD + BPD (α, Θ)	88.8	84.4	82.8	80.2	81.4
DD + BPD (α, ϕ)	92.1	92.6	89.5	88.8	88.1
DD + BPD (ϕ)	92.3	91.6	90	88.4	87

To analyze which actions are most often misclassified we calculated the confusion matrix for the best configuration: DD + BPD (α, ϕ). The matrix is presented in Figure 3.

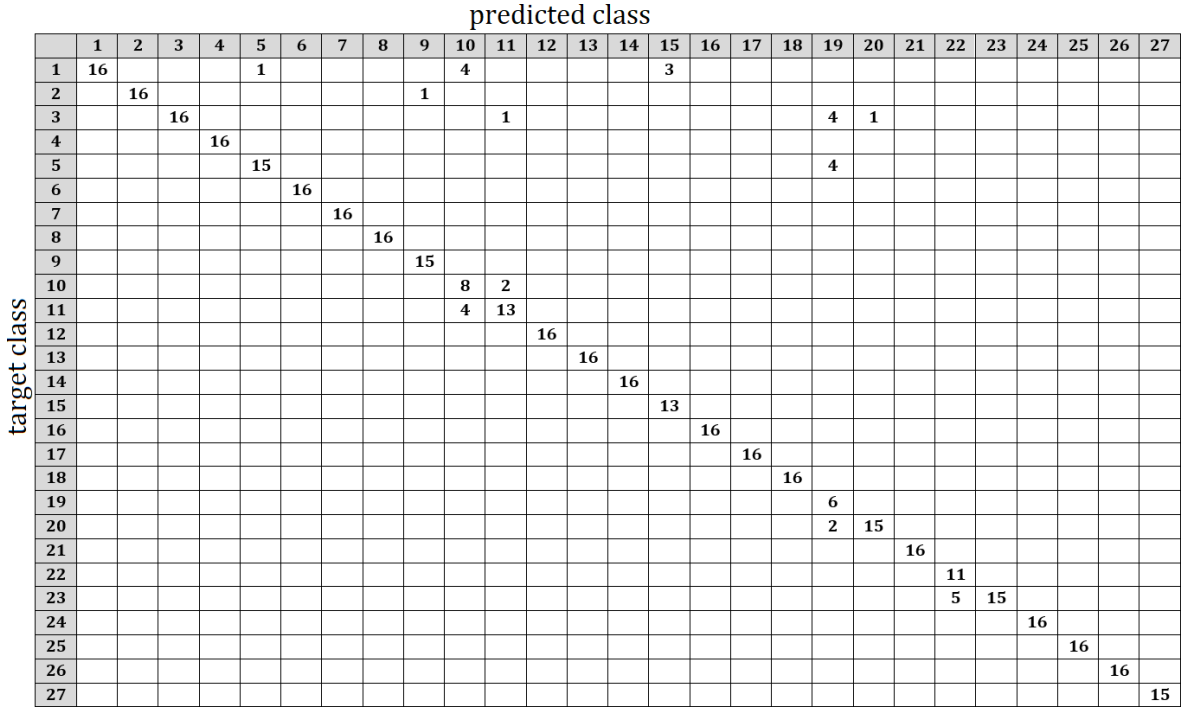

Figure 3. The confusion matrix of the proposed method for the best configuration of features and parameters: DD + BPD (α, ϕ), $K = 2$. The actions corresponding to particular numbers are listed in Section 4.1.

As one can see, many actions were recognized perfectly. There is only one action, "right hand knock on door", with recognition rate below 50%. It was confused with "right hand wave" and

"right arm throw" four times and with "right hand catch an object" twice. Another common misclassification is confusing "jogging in place" with "walking in place". Moreover, "right hand draw circle (counterclockwise)" was confused with "right arm swipe to the left" and "draw triangle" four times. All of these mistakes occur for the actions that, in the case of some subjects, are difficult to differentiate even for a human. The actions that are visually much different from each other were recognized with almost 100% accuracy.

In Table 3, we present the comparison of our approach with other existing methods which use only skeletal data from UTD-MHAD dataset. The proposed algorithm outperforms five of the listed methods almost equaling with the best found method, Bayesian Hierarchical Dynamic Model (HDM) [6].

Table 3. Comparison of the proposed method with other existing methods using UTD-MHAD dataset.

Method	Recognition Rate [%]
Label Consistent K-SVD [2,4]	76.2
Covariance Joint Descriptors [1,4]	85.6
Optical Spectra-based CNN [34]	87
Joint Trajectory Maps [4]	87.9
Joint Distance Maps [5]	88.1
Our method (DD + BPD)	92.6
Bayesian HDM [6]	92.8

The average time of extracting features from an action (using the best configuration: DD + BPD (α, ϕ)) is ~600 ms and the average classification time using LDMLT is ~300 ms. Therefore, the average recognition time of a single action is below 1 s. The total training time is about 550 s. Most of the works presenting methods that use only skeletal data (see Table 3) do not report computational times of their algorithms. The only exception is Hou et al. [34], whose method Optical Spectra-based CNN achieved very short recognition time of about 40 ms on a PC with Intel Core i7-4790, 4 GHz CPU and NVIDIA TITAN X GPU (some parts of code were running on GPU). However, the total training time of this method is about 880 s, which is relatively long. It is also worth noting that Optical Spectra-based CNN achieved recognition rate of 5.6 percentage points less than our method despite being much faster in terms of recognition time.

5. Conclusions

In this paper, we proposed recognition method for human actions using skeletal data. Two descriptors, originally intended for hand posture recognition, were successfully adapted for classification of the body skeleton sequences. One of these descriptors, BPD, was modified to replace surface normals and vectors pointed by fingers with vectors representing bones. Configurations of joints and bones, used as an input data for descriptors, were selected. The experimental tests were performed using five classifiers and configurations of features. Our method achieved high recognition rate compared to other existing methods, which confirms its usefulness.

The proposed method does not require a specific lighting conditions, background, or any special outfit, e.g., inertial sensors and gloves. Moreover, it is fast enough to run in real-time. However, the implementation can be optimized in terms of recognition time, which can be a subject for a future work. The future work may also include combining skeletal data with depth maps/point clouds and color/gray images to develop more effective feature vectors. Such a combination could improve distinguishing between actions like "knock on door", "hand wave", and "arm throw" as features based on point clouds and color images, unlike skeletal descriptors, are able to recognize hand shapes. Another future study topic may be expansion of training datasets by generating additional action sequences based on the existing data.

Author Contributions: Conceptualization and methodology, D.W. and T.K.; software, D.W. and T.K.; validation, experiments design, and discussion of the results, D.W.; writing—original draft preparation, and review and editing, D.W. and T.K. All authors have read and agreed to the published version of the manuscript.

Funding: This project is financed by the Minister of Science and Higher Education of the Republic of Poland within the "Regional Initiative of Excellence" program for years 2019–2022. Project number: 027/RID/2018/19, amount granted: 11 999 900 PLN.

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

References

1. Hussein, M.E.; Torki, M.; Gowayyed, M.A.; El-Saban, M. Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*; AAAI Press: Beijing, China, 2013; pp. 2466–2472.
2. Zhou, L.; Li, W.; Zhang, Y.; Ogunbona, P.; Nguyen, D.T.; Zhang, H. Discriminative Key Pose Extraction Using Extended LC-KSVD for Action Recognition. In Proceedings of the 2014 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Wollongong, NSW, Australia, 25–27 November 2014; pp. 1–8.
3. Chen, C.; Jafari, R.; Kehtarnavaz, N. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015.
4. Wang, P.; Li, Z.; Hou, Y.; Li, W. Action Recognition Based on Joint Trajectory Maps Using Convolutional Neural Networks. In *Proceedings of the 24th ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2016.
5. Li, C.; Hou, Y.; Wang, P.; Li, W. Joint Distance Maps Based Action Recognition With Convolutional Neural Networks. *IEEE Signal Process. Lett.* **2017**, *24*, 624–628. [CrossRef]
6. Zhao, R.; Xu, W.; Su, H.; Ji, Q. Bayesian Hierarchical Dynamic Model for Human Action Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7733–7742.
7. Yang, X.; Zhang, C.; Tian, Y. Recognizing actions using depth motion maps-based histograms of oriented gradients. In Proceedings of the 20th ACM International Conference on Multimedia, Nara, Japan, 27–31 October 2012; pp. 1057–1060.
8. Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time human action recognition based on depth motion maps. *J. Real-Time Image Process.* **2016**, *12*, 155–163. [CrossRef]
9. Li, W.; Zhang, Z.; Liu, Z. Action recognition based on a bag of 3D points. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Workshops, San Francisco, CA, USA, 13-18 June 2010; pp. 9–14.
10. Bulbul, M.F.; Jiang, Y.; Ma, J. Human action recognition based on DMMs, HOGs and Contourlet transform. In Proceedings of the 2015 IEEE International Conference on Multimedia Big Data, Beijing, China, 20–22 April 2015; pp. 389–394.
11. Chen, C.; Liu, M.; Liu, H.; Zhang, B.; Han, J.; Kehtarnavaz, N. Multi-temporal depth motion maps-based local binary patterns for 3-D human action recognition. *IEEE Access* **2017**, *5*, 22590–22604. [CrossRef]
12. Zhang, B.; Yang, Y.; Chen, C.; Yang, L.; Han, J.; Shao, L. Action recognition using 3D histograms of texture and a multi-class boosting classifier. *IEEE Trans. Image Process.* **2017**, *26*, 4648–4660. [CrossRef]
13. Yang, X.; Tian, Y. Super normal vector for activity recognition using depth sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 804–811.
14. Slama, R.; Wannous, H.; Daoudi, M. Grassmannian representation of motion depth for 3D human gesture and action recognition. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 3499–3504.
15. Liu, M.; Liu, H. Depth context: A new descriptor for human activity recognition by using sole depth sequences. *Neurocomputing* **2016**, *175*, 747–758. [CrossRef]
16. Liu, M.; Liu, H.; Chen, C. Robust 3D action recognition through sampling local appearances and global distributions. *IEEE Trans. Multimed.* **2017**, *20*, 1932–1947. [CrossRef]

17. Liu, B.; Cai, H.; Ju, Z.; Liu, H. RGB-D sensing based human action and interaction analysis: A survey. *Pattern Recognit.* **2019**, *94*, 1–12. [CrossRef]
18. Qiao, R.; Liu, L.; Shen, C.; van den Hengel, A. Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition. *Pattern Recognit.* **2017**, *66*, 202–212. [CrossRef]
19. Devanne, M.; Wannous, H.; Berretti, S.; Pala, P.; Daoudi, M.; Del Bimbo, A. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE Trans. Cybern.* **2014**, *45*, 1340–1352. [CrossRef]
20. 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]
21. Lillo, I.; Niebles, J.C.; Soto, A. Sparse composition of body poses and atomic actions for human activity recognition in RGB-D videos. *Image Vis. Comput.* **2017**, *59*, 63–75. [CrossRef]
22. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Learning actionlet ensemble for 3D human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 914–927. [CrossRef] [PubMed]
23. Raman, N.; Maybank, S.J. Activity recognition using a supervised non-parametric hierarchical HMM. *Neurocomputing* **2016**, *199*, 163–177. [CrossRef]
24. Shahroudy, A.; Ng, T.T.; Yang, Q.; Wang, G. Multimodal multipart learning for action recognition in depth videos. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 2123–2129. [CrossRef] [PubMed]
25. Sung, J.; Ponce, C.; Selman, B.; Saxena, A. Unstructured human activity detection from rgbd images. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012; pp. 842–849.
26. Zhang, Z.; Ma, X.; Song, R.; Rong, X.; Tian, X.; Tian, G.; Li, Y. Deep learning based human action recognition: A survey. In Proceedings of the 2017 Chinese Automation Congress (CAC), Jinan, China, 20–22 October 2017; pp. 3780–3785.
27. Du, Y.; Fu, Y.; Wang, L. Skeleton based action recognition with convolutional neural network. In Proceedings of the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, Malaysia, 3–6 November 2015; pp. 579–583.
28. Li, B.; Dai, Y.; Cheng, X.; Chen, H.; Lin, Y.; He, M. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017; pp. 601–604.
29. Li, C.; Zhong, Q.; Xie, D.; Pu, S. Skeleton-based action recognition with convolutional neural networks. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017; pp. 597–600.
30. Ke, Q.; An, S.; Bennamoun, M.; Sohel, F.; Boussaid, F. Skeletonnet: Mining deep part features for 3-d action recognition. *IEEE Signal Process. Lett.* **2017**, *24*, 731–735. [CrossRef]
31. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. A new representation of skeleton sequences for 3d action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3288–3297.
32. Ding, Z.; Wang, P.; Ogunbona, P.O.; Li, W. Investigation of different skeleton features for cnn-based 3d action recognition. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, China, 10–14 July 2017; pp. 617–622.
33. Imran, J.; Kumar, P. Human action recognition using RGB-D sensor and deep convolutional neural networks. In Proceedings of the 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 21–24 September 2016; pp. 144–148.
34. Hou, Y.; Li, Z.; Wang, P.; Li, W. Skeleton Optical Spectra-Based Action Recognition Using Convolutional Neural Networks. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 807–811. [CrossRef]
35. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.
36. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.

37. Wang, H.; Wang, L. Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3633–3642.
38. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 816–833.
39. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Li, F.-F. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
40. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
41. Li, Y.; Li, W.; Mahadevan, V.; Vasconcelos, N. Vlad3: Encoding dynamics of deep features for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1951–1960.
42. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
43. Singh, B.; Marks, T.K.; Jones, M.; Tuzel, O.; Shao, M. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1961–1970.
44. Mahasseni, B.; Todorovic, S. Regularizing long short term memory with 3D human-skeleton sequences for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3054–3062.
45. Xin, M.; Zhang, H.; Wang, H.; Sun, M.; Yuan, D. ARCH: Adaptive recurrent-convolutional hybrid networks for long-term action recognition. *Neurocomputing* **2016**, *178*, 87–102. [CrossRef]
46. Xin, M.; Zhang, H.; Sun, M.; Yuan, D. Recurrent Temporal Sparse Autoencoder for attention-based action recognition. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 456–463.
47. Donahue, J.; Hendricks, L.A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 677–691. [CrossRef]
48. Aggarwal, J.; Ryoo, M. Human Activity Analysis: A Review. *ACM Comput. Surv.* **2011**, *43*. [CrossRef]
49. Kapuściński, T.; Warchoł, D. Hand Posture Recognition Using Skeletal Data and Distance Descriptor. *Appl. Sci.* **2020**, *10*, 2132. [CrossRef]
50. Kapuściński, T.; Organiściak, P. Handshape Recognition Using Skeletal Data. *Sensors* **2018**, *18*, 2577. [CrossRef] [PubMed]
51. Rusu, R.B.; Marton, Z.C.; Blodow, N.; Beetz, M. Learning informative point classes for the acquisition of object model maps. In Proceedings of the 2008 10th International Conference on Control, Automation, Robotics and Vision, Hanoi, Vietnam, 17–20 December 2008; pp. 643–650.
52. Spivak, M. *A Comprehensive Introduction to Differential Geometry*, 3rd ed.; Publish or Perish: Houston, TX, USA, 1999; Volume 3.
53. Matlab Scripts for Distance Descriptor and Bone Pair Descriptor. Available online: http://vision.kia.prz.edu.pl (accessed on 9 February 2020).
54. Celebi, S.; Aydin, A.S.; Temiz, T.T.; Arici, T. Gesture Recognition using Skeleton Data with Weighted Dynamic Time Warping. In Proceedings of the International Conference on Computer Vision Theory and Applications—VISAPP 2013, Barcelona, Spain, 21–24 February 2013; pp. 620–625.
55. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
56. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.

57. Graves, A.; Jaitly, N.; Mohamed, A. Hybrid speech recognition with Deep Bidirectional LSTM. In Proceedings of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, 8–12 December 2013; pp. 273–278.
58. Aceto, G.; Ciuonzo, D.; Montieri, A.; Pescape, A. Mobile Encrypted Traffic Classification Using Deep Learning: Experimental Evaluation, Lessons Learned, and Challenges. *IEEE Trans. Netw. Serv. Manag.* **2019**, *16*, 445–458. [CrossRef]
59. Mei, J.; Liu, M.; Karimi, H.R.; Gao, H. LogDet Divergence-Based Metric Learning with Triplet Constraints and Its Applications. *IEEE Trans. Image Process.* **2014**, *23*, 4920–4931. [CrossRef]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).