A discussion on the validation tests employed to compare human action recognition methods using the MSR Action3D dataset

José Ramón Padilla López[1], Alexandros André Chaaraoui[1], and Francisco Flórez Revuelta[2]

[1] Department of Computer Technology, University of Alicante,
P.O. Box 99, E-03080 Alicante, Spain
jpadilla@dtic.ua.es, alexandros@dtic.ua.es
[2] Faculty of Science, Engineering and Computing, Kingston University,
Penrhyn Road, KT1 2EE, Kingston upon Thames, United Kingdom
F.Florez@kingston.ac.uk

Abstract. This paper aims to determine which is the best human action recognition method based on features extracted from RGB-D devices, such as the Microsoft Kinect. A review of all the papers that make reference to MSR Action3D, the most used dataset that includes depth information acquired from a RGB-D device, has been performed. We found that the validation method used by each work differs from the others. So, a direct comparison among works cannot be made. However, almost all the works present their results comparing them without taking into account this issue. Therefore, we present different rankings according to the methodology used for the validation in orden to clarify the existing confusion.

Keywords: human action recognition, RGB-D devices, MSR Action3D, validation, Kinect, depth sensors

1 Introduction

In recent years, interest has grown on affordable devices (*e.g. Microsoft Kinect* or *ASUS Xtion Pro*) that capture depth quite reliably. Such devices provide a depth image (D), along with an RGB image (thus RGB-D). A depth image can be further processed to obtain marker-less body pose estimation by means of a skeleton model consisting of a series of joints. Due to their low cost, high sample rate and capability to combine visual and depth information, these devices have become widespread in both research and commercial applications. Furthermore, their use has not been restricted to games, for which they were initially designed, but other applications where natural human-computer interaction is required.

These devices are widely used in the field of human action recognition (HAR), particularly in indoor scenarios for the recognition of activities of daily living. For research purposes, a variety of datasets for human action (or gesture) recognition have been recorded using RGB-D devices (see Table 1). The MSR Action3D

Table 1. State-of-the-art datasets for action recognition based on depth or skeletal features, sorted from more quoted to less quoted according to Google Scholar.

Name	Actions	Actors	Times	Samples	Citations	Year
MSR Action3D [1]	20	10	2 or 3	567	176	2010
MSR DailyActivity3D [2]	16	10	2	320	138	2012
RGBD-HuDaAct [3]	12	30	2 or 4	1189	86	2011
CAD-60 [4]	12	2+2	-	60	80	2012
UTKinect Action [5]	10	10	2	-	73	2012
MSRC-12 KinectGesture [6]	12	30	-	594	39	2012
CAD-120 [7]	10	2+2	-	120	33	2013
MSR ActionPairs [8]	6	10	3	180	29	2013
MSR Gesture3D [9]	12	10	2 or 3	336	25	2012
LIRIS Human Activities [10]	10	21	-	-	24	2012
Berkeley MHAD [11]	11	7+5	5	~ 660	18	2013
G3D [12]	20	10	3	-	11	2012
ACT4 Dataset [13]	14	24	>1	6844	9	2012
UPCV Action [14]	10	20	-	-	6	2014
WorkoutSu-10 Gesture [15]	10	15	10	1500	6	2013
IAS-Lab Action [16]	15	12	3	540	3	2013
Florence 3D Action [17]	9	10	2 or 3	215	2	2012

dataset [1] from Microsoft Research stands out as one of the most used in the literature, as many developed methods for action recognition have been validated with this dataset. Hence, it should be easy to determine the best human action recognition method in a straightforward way by comparing their success and processing rates. However, to the best of our knowledge, this is not possible at the moment as we found that almost all the works compare results obtained with different validation methods.

Therefore, this work aims to fill the existing gap in order to enable a fair comparison of the state of the art. We have reviewed 176 papers that make reference to the MSR Action3D dataset. Out of these 176 papers, 62 papers have been considered as they use the MSR Action3D dataset for the validation of the human action (or gesture) recognition methods proposed. They are classified according to the validation method and ranked based on their success rate.

The remainder of this paper is organised as follows: Section 2 describes the MSR Action3D dataset employed by the reviewed works. In section 3, an explanation of the inconsistencies found in the number of the used samples is given. Section 4 presents the validation methods used in the reviewed papers and provides a classification of each work according to this. Finally, section 5 presents some conclusions and recommendations for the future.

2 MSR Action3D dataset

The MSR Action3D dataset [1] contains 20 different actions, performed by 10 different subjects with up to 3 different repetitions. This makes a total of 567

Table 2. Actions in each of the MSR Action3D subsets.

AS1		AS2		AS3	
Label	Action name	Label	Action name	Label	Action name
a02	Horizontal arm wave	a01	High arm wave	a06	High throw
a03	Hammer	a04	Hand catch	a14	Forward kick
a05	Forward punch	a07	Draw cross	a15	Side-kick
a06	High throw	a08	Draw tick	a16	Jogging
a10	Hand clap	a09	Draw circle	a17	Tennis swing
a13	Bend	a11	Two-hand wave	a18	Tennis serve
a18	Tennis serve	a14	Forward kick	a19	Golf swing
a20	Pick-up and throw	a12	Side-boxing	a20	Pick-up and throw

sequences and each one includes depth and skeleton joints. However 10 sequences are not valid in this dataset because the skeletons were either missing or wrong, as explained by the authors[3]. The authors divided the dataset in three subsets of 8 gestures each, as shown in Table 2. Most of the papers working with this dataset have also used them. This was due to the high computational cost of dealing with the overall dataset. The AS1 and AS2 subsets were intended to group actions with similar movement, while AS3 was intended to group complex actions together.

3 How many samples are used for testing?

Despite of the fact that the MSR Action3D dataset is made up of 567 sequences, the number of instances used in some works is unclear [18–20]. There is a lot of confusion concerning this topic.

As far as we know, the authors of the dataset firstly described it as made up of twenty actions, where each one was performed by seven subjects for three times [1]. However, actions are performed by ten subjects with up to three repetitions as described in the previous section. Many works have compared their results with Li et al. and most of them used ten subjects [5, 21, 22]. In other words, they may have used a higher number of instances than the work they aim to compare to. Wang et al. [2] described the dataset as made up of 402 sequences. For the sake of clarity, this mistake is advertised at the dataset web page[4]. The authors explain that 10 sequences out of the 567 are not used because a number of skeletons are either missing or too erroneous. So, the dataset is eventually composed of 557 sequences. However, it is curious to see how recent works [19, 20, 23] still mention that the dataset is composed of 402 sequences and directly compare their results with the state-of-the-art papers that use other number of instances. Furthermore, other authors have intentionally used a subset of the whole dataset, *e.g.* 17 actions, 8 subjects and 3 repetitions (408 samples). Due

[3] MSR Action Recognition Datasets and Codes, http://research.microsoft.com/en-us/um/people/zliu/actionrecorsrc/default.htm (last access: 06/26/2014)

[4] A list of the used sequences is also provided in the website

to this, the AS1, AS2 and AS3 subsets are composed of different actions too, thereby they compare their results with works that use a different number of instances.

As a consequence, it is very difficult to confirm whether these works use 402, 557 or 567 samples as we are not sure whether the authors are aware of these key aspects concerning the dataset, or if those are only naive text mistakes. Moreover, the missing information concerning the number of instances prevents to make a fair comparison between different methods.

4 Which is the validation method used?

Regarding the experimentation method used by many authors working with the MSR Action3D dataset, it is worth to mention that there is a lack of agreement. In the paper by Li et al. [1] where the dataset was firstly presented, three tests are performed: 1/3, 2/3 and cross-subject test. In the first two tests, 1/3 and 2/3 of the instances are respectively used as training samples and the rest as testing samples. In the third test, half of the subjects are used for training and the remainder for testing. However, it is not described which instances or subjects are actually used in each partition of the dataset.

Given that information is missing, we could assume that the 1/3 means to split the dataset using the first repetition of each action performed by each subject as training, and to use the remainder for testing. The same could be assumed for the 2/3. However, if we only consider instances as a whole, we can split the dataset in a different way. For instance, the dataset can be split using 1/3 (or 2/3) of all the instances for training. The same is true for the cross-subject test. It is not stated which instances are used. Any half of all the subjects can be used for training, *e.g.* 1, 2, 3, 9 and 10; and the remainder for testing, *i.e.* 4, 5, 6, 7 and 8. Given that it is not clear which instances are used, each researcher is free to interpret anything, thereby comparing different methods where a distinct methodology has been used for the experimentation. However, this is not desirable to compare and decide which method performs better.

In the cross-subject test employed by Li et al. [1] the actual samples of subjects 1, 3, 5, 7 and 9 are used for training, whereas actors 2, 4, 6, 8 and 10 are used for validation. This test is followed by many authors as shown in Table 3. While some authors use the mentioned settings for their training and validation sets, other authors use subjects 1-5 for training and 6-10 for validation (see Table 4). Regardless of the used setup, most of the works state that they follow the same settings as Li et al. but do not provide a description of such a setup. Due to this, we assume that they follow the same validation than Li et al., so Table 3 and Table 4 can even have classification mistakes. Anyway, a fair comparison cannot be performed. Indeed, when it is sure that the same setup has been used, sometimes results only show an accuracy score and the authors do not give an explanation of what it represents, *i.e.* the average of the AS1, AS2 and AS3 tests, or the overall accuracy of using the whole dataset (20 actions).

Table 3: Li et al.'s cross-subject test. The first eight methods explicitly describe both the training and validation sets. Results are ordered by the average result for the AS1, AS2, and AS3 subsets; and then by the results for the whole dataset.

Method	**Year**	**AS1**	**AS2**	**AS3**	**Avg.**	**All**
Fusion of Skeletal and Silhouette-Based Features for Human Action Recognition with RGB-D Devices, Chaaraoui et al. [24]	2013	92.38	86.61	96.4	91.8	-
Real-time human action recognition based on depth motion maps, Chen et al. [25]	2013	96.2	83.2	92	90.47	-
Skeletal Quads: Human Action Recognition Using Joint Quadruples, Evangelidis et al. [26]	2014	88.39	86.61	94.59	89.86	-
Random Occupancy Patterns, Wang et al. [27]	2014	-	-	-	86.50?	-
Action recognition based on a bag of 3d points, Li et al. [1]	2010	72.9	71.9	79.2	74.67	-
Learning Maximum Margin Temporal Warping for Action Recognition, Wang and Wu [21]	2013	-	-	-	-	92.7?
Learning Actionlet Ensemble for 3D Human Action Recognition, Yuan et al. [23]	2014	-	-	-	-	88.2?
Mining actionlet ensemble for action recognition with depth cameras, Wang et al. [2]	2012	-	-	-	-	88.2?
Group Sparsity and Geometry Constrained Dictionary Learning for Action Recognition from Depth Maps, Luo et al. [28]	2013	97.2	95.5	99.1	97.26	96.7
Fusing Spatiotemporal Features and Joints for 3D Action Recognition, Zhu et al. [29]	2013	-	-	-	94.3	-
Human Action Recognition by Mining Discriminative Segment with Novel Skeleton Joint Feature, Zou et al. [30]	2013	-	-	-	94.0?	
Pose-based human action recognition via sparse representation in dissimilarity space , Theodorakopoulos et al. [14]	2014	91.23	90.09	99.5	93.61	-
Action recognition on motion capture data using a dynemes and forward differences representation, Kapsouras and Nikolaidis [31]	2014	-	-	-	93.6	91.4
Super Normal Vector for Activity Recognition Using Depth Sequences, Yang and Tian [32]	2014	-	-	-	93.09?	-
Action Recognition Using Ensemble Weighted Multi-Instance Learning, Chen et al. [33]	2014	-	-	-	92?	-
Histogram of Oriented Displacements (HOD): Describing Trajectories of Human Joints for Action Recognition, Gowayyed et al. [34]	2013	92.39	90.18	91.43	91.26	-
Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations, Hussein et al. [35]	2013	88.04	89.29	94.29	90.53	-
Body Surface Context: A New Robust Feature for Action Recognition From Depth Videos, Song et al. [36]	2014	-	-	-	90.36?	-
An Approach to Pose-Based Action Recognition, Wang et al. [37]	2013	-	-	-	90.22	-
Human Action Recognition Via Multi-modality Information, Gao et al. [19]	2014	92	85	93	90	-

Continued on next page

Table 3 – *Continued from previous page*

Method	Year	AS1	AS2	AS3	Avg.	All
Human Behavior Recognition Based on Axonometric Projections and PHOG Feature, Shen et al. [20]	2014	90.6	81.4	94.6	88.87	-
On the improvement of human action recognition from depth map sequences using Space-Time Occupancy Patterns, Vieira et al. [38]	2014	91.7	72.2	98.6	87.5	81.55
STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences, Vieira et al. [39]	2012	84.7	81.3	88.4	84.8	-
Effective 3D action recognition using EigenJoints, Yang And Tian [22]	2014	-	-	-	83.3?	-
Home Monitoring Musculo-skeletal Disorders with a Single 3D Sensor, Wang et al. [40]	2013	-	-	-	81.9?	-
Online Human Gesture Recognition from Motion Data Streams, Zhao et al. [41]	2013	-	-	-	81.7?	-
Effective approaches in human action recognition, Li et al. [42]	2013	-	-	-	81.5 or 91.5?	-
Gesture recognition from depth images using motion and shape features, Qin et al. [43]	2013	81	79	82	80.66	-
Human activity recognition using multi-features and multiple kernel learning, Althloothi et al. [44]	2014	74.3	76.8	86.7	79.27	-
View invariant human action recognition using histograms of 3D joints, Xia et al. [5]	2012	87.98	85.48	63.46	78.97	-
Three Dimensional Motion Trail Model for Gesture Recognition, Liang and Zheng [45]	2013	73.7	81.5	81.6	78.93	-
Attractor-Shape for Dynamical Analysis of Human Movement: Applications in Stroke Rehabilitation and Action Recognition, Venkataraman et al. [46]	2013	77.5	63.1	87	75.87	-
Exploring the Trade-off Between Accuracy and Observational Latency in Action Recognition, Ellis et al. [47]	2013	-	-	-	65.7?	-
Robust 3D Action Recognition with Random Occupancy Patterns, Wang et al. [48]	2012	-	-	-	-	86.5?

Table 4: Cross-subject test (1-5 training, 6-10 test). The first seven methods explicitly describe both the training and validation sets.

Method	Year	AS1	AS2	AS3	Avg.	All
Sparse spatio-temporal representation of joint shape-motion cues for human action recognition in depth sequences, Tran and Ly [49]	2013	-	-	-	91.92?	-
The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection, Zanfir et al. [50]	2013	-	-	-	91.7?	-
An effective fusion scheme of spatio-temporal features for human action recognition in RGB-D video, Tran and Ly [51]	2013	-	-	-	88.89?	-

Continued on next page

Table 4 – *Continued from previous page*

Method	Year	AS1	AS2	AS3	Avg.	All
Real Time Action Recognition Using Histograms of Depth Gradients and Random Decision Forests, Rahmani et al. [52]	2014	-	-	-	88.8?	-
Iterative temporal learning and prediction with the sparse online echo state gaussian process, Soh and Demiris [53]	2012	80.6	74.9	87.1	80.87	-
Joint Angles Similarities and HOG2 for Action Recognition, Ohn-Bar and Trivedi [54]	2013	-	-	-	-	94.84
HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences, Oreifej and Liu [8]	2013	-	-	-	-	88.89?
Spatio-temporal feature extraction and representation for RGB-D human action recognition, Luo et al. [55]	2014	96.1	90.8	98.33	95.08	93.83?
Action Classification with Locality-constrained Linear Coding, Rahmani et al. [56]	2014	-	-	-	90.9?	-
Spatio-temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera, Xia and Aggarwal [57]	2013	-	-	-	89.3?	-
Optimal Joint Selection for Skeletal Data from RGB-D Devices Using a Genetic Algorithm, Climent et al. [58]	2013	-	-	-	-	71.1

Due to all this confusion about how to split the dataset in two sets for training and validation, some authors randomly choose half of the subjects for the training set, and select the rest of the subjects for the validation set. As in 2-fold cross validation, they repeat the test using the previous validation set as the training set and vice versa. In this case, the final result is the average of both tests (see Table 5). In other works, instead of performing a 2-fold cross validation, some authors randomly select the two sets and repeat the experiment several times. For example, Miranda et al. [59] perform a random selection of half of the actors as training set and the other half as validation set. This is repeated 10 times and the final result is the average of the results of each run. Other authors repeat the test 100 times instead of 10 [60, 61], and even 200 times [25] (see Table 6). However, although the tests are repeated many times, all the possible splits are not considered, *i.e.* all the possible combinations (252 tests) of using 5 subjects for training and the remaining ones for testing. Only three works perform this test [8, 49, 52]. In Table 6 these works have been included with the 252 number in the third column. This indicates that they performed a test with all the possible combinations.

Another approach used by some authors is to perform a leave-one-actor-out cross-validation test. In this case, actor invariance is specifically tested by training with all but one actor, and testing the method with the unseen one. This is repeated for all the actors, averaging the returned success rates (see Table 7).

Finally, in addition to the described validation methods that are frequently used in the literature, there are other authors that have not been included in any table because either the validation method is unclear [66, 67] or the employed

Table 5. 2-fold cross-validation test

Method	Year	AS1	AS2	AS3	Avg.	All
Evolutionary joint selection to improve human action recognition with RGB-D devices, Chaaraoui et al. [62]	2014	91.59	90.83	97.28	93.23	-
3D Action Classification Using Sparse Spatio-temporal Feature Representations, Azary and Savakis [63]	2012	77.66	73.17	91.58	80.8	63.23

Table 6. Miranda et al.'s test: Random selection of training and test sets repeated a number of times.

Method	Year	Tests	AS1	AS2	AS3	Avg.	All
Sparse spatio-temporal representation of joint shape-motion cues for human action recognition in depth sequences, Tran and Ly [49]	2013	252	-	-	-	84.54?	-
Real Time Action Recognition Using Histograms of Depth Gradients and Random Decision Forests, Rahmani et al. [52]	2014	252	-	-	-	-	82.7?
HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences, Oreifej and Liu [8]	2013	252	-	-	-	-	82.15?
Real-time human action recognition based on depth motion maps, Chen et al. [25]	2013	200	90.1	90.6	97.6	92.77	-
Fast Exact Hyper-graph Matching with Dynamic Programming for Spatio-temporal Data, Çeliktutan et al. [61]	2014	100	84.5	85	72.2	80.57	-
Graph-based Analysis of Physical Exercise Actions, eliktutan et al. [60]	2013	100	84.5	85	72.2	80.5	-
Online gesture recognition from pose kernel learning and decision forests, Miranda et al. [64]	2014	10	96	57.1	97.3	83.5	-
Real-Time Gesture Recognition from Depth Data through Key Poses Learning and Decision Forests, Miranda et al. [59]	2012	10	93.5	52	95.4	80.3	-
Space-Time Pose Representation for 3D Human Action Recognition, Devanne et al. [65]	2013	10	84.8	67.8	87.1	79.9	-

Table 7. Leave-one-actor-out cross-validation test

Method	Year	AS1	AS2	AS3	Avg.	All
Evolutionary joint selection to improve human action recognition with RGB-D devices, Chaaraoui et al. [62]	2014	91.46	91.78	97.13	93.46	-
Fusion of Skeletal and Silhouette-Based Features for Human Action Recognition with RGB-D Devices, Chaaraoui et al. [24]	2013	90.65	85.15	95.93	90.58	-
3D Action Classification Using Sparse Spatio-temporal Feature Representations, Azary and Savakis [63]	2012	80.73	77.11	93.89	83.91	72.11
Grassmannian Sparse Representations and Motion Depth Surfaces for 3D Action Recognition, Azary and Savakis [18]	2013	-	-	-	-	78.48?
Fast Exact Hyper-graph Matching with Dynamic Programming for Spatio-temporal Data, Çeliktutan et al. [61]	2014	-	-	-	-	72.9?

settings are not used by more than one author [68–71]. For instance, Ofli et al. [68] use a subset of 17 actions and 8 subjects. They train with 5 subjects and validate with 3 subjects in order to obtain the success rate (41.18% for the whole dataset). These results are improved in [70] with the same setup (83.89% for the whole dataset). Cottone et al. [69] perform a leave-one-sequence-out cross validation, training with all the sequences in the dataset but one that is used for testing. Then, they perform 10 of these tests obtaining the average success rate (90.47% for the average of AS1, AS2 and AS3). Sabinas et al.[71] are focused on early detection of gestures, *i.e.* without seeing all the information, and their experimentation is based on one-shot learning. Therefore, their results are not directly comparable (47% for the average of AS1, AS2 and AS3).

5 Conclusions

In this work, we have aimed to give an answer to the question of which is the best action recognition method based on features extracted from depth and skeletal data. Based on the review the present work has performed, it can be observed that we cannot answer this question with confidence. In other words, we cannot know so far. Hence, we have presented the most important divergences in the comparison of action recognition methods that use the MSR Action3D dataset. Among these, we can highlight the mismatch in the number of samples used by most of the works and the different validation methods that have been used. As we have seen, the validation performed by Li et al. is one of the most used. However, the missing information about how to split the dataset into training and validation sets has led to a lot of confusion. Furthermore, most of the authors do not describe how this division is performed in their works. Therefore, experiments cannot be reproduced and fair comparisons cannot be made. Thus, in this work we have tried to clear up the existing confusion. This may enable to improve future comparisons and increase the awareness of the need of clarifying experimental settings.

Among all the validation methods reviewed in this work, we consider that the cross validation considering all the possible splits of the dataset, *i.e.* all the possible combinations (252 tests) of using 5 subjects for training and the remaining ones for testing, is the most robust validation method. However, if testing your method with the 5-5 splits cross validation is very demanding concerning computational cost, then the leave-one-actor-out cross validation is the one we recommend under these conditions.

Notes to authors

As it has been difficult in some cases to understand the validation method of the papers, we encourage authors of the reviewed works to contact us in case their works had been misclassified in the previous tables. This way, we will be able to update the document and correct it. Similarly, authors of new works are also encouraged to contact us in order to incorporate their works if so desired.

References

1. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2010) 9–14
2. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. (June 2012) 1290–1297
3. Ni, B., Wang, G., Moulin, P.: RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). (Nov 2011) 1147–1153
4. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from RGBD images. In: 2012 IEEE International Conference on Robotics and Automation (ICRA). (May 2012) 842–849
5. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3D joints. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2012) 20–27
6. Fothergill, S., Mentis, H.M., Kohli, P., Nowozin, S.: Instructing people for training gestural interactive systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM (2012) 1737–1746
7. Koppula, H.S., Gupta, R., Saxena, A.: Learning Human Activities and Object Affordances from RGB-D Videos. International Journal of Robotics Research **32**(8) (2013) 951–970
8. Oreifej, O., Liu, Z.: HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2013) 716–723
9. Kurakin, A., Zhang, Z., Liu, Z.: A real time system for dynamic hand gesture recognition with a depth sensor. In: Proceedings of the 20th European Signal Processing Conference (EUSIPCO). (Aug 2012) 1975–1979
10. Wolf, C., Mille, J., Lombardi, E., Celiktutan, O., Jiu, M., Baccouche, M., Dellandréa, E., Bichot, C.E., Garcia, C., Sankur, B.: The LIRIS Human activities dataset and the ICPR 2012 human activities recognition and localization competition. Technical Report RR-LIRIS-2012-004, LIRIS Laboratory (March 2012)
11. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: Berkeley MHAD: A comprehensive Multimodal Human Action Database. In: 2013 IEEE Workshop on Applications of Computer Vision (WACV). (Jan 2013) 53–60
12. Bloom, V., Makris, D., Argyriou, V.: G3D: A gaming action dataset and real time action recognition evaluation framework. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2012) 7–12
13. Cheng, Z., Qin, L., Ye, Y., Huang, Q., Tian, Q.: Human Daily Action Analysis with Multi-view and Color-Depth Data. In: Computer Vision - ECCV 2012. Workshops and Demonstrations. Volume 7584 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 52–61
14. Theodorakopoulos, I., Kastaniotis, D., Economou, G., Fotopoulos, S.: Pose-based human action recognition via sparse representation in dissimilarity space. Journal of Visual Communication and Image Representation **25**(1) (2014) 12–23
15. Negin, F., Özdemir, F., Akgül, C., Yüksel, K.A., Erçil, A.: A Decision Forest Based Feature Selection Framework for Action Recognition from RGB-Depth Cameras. In: Image Analysis and Recognition. Volume 7950 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 648–657

16. Munaro, M., Ballin, G., Michieletto, S., Menegatti, E.: 3D flow estimation for human action recognition from colored point clouds. Biologically Inspired Cognitive Architectures **5** (2013) 42–51
17. Seidenari, L., Varano, V., Berretti, S., Del Bimbo, A., Pala, P.: Recognizing Actions from Depth Cameras as Weakly Aligned Multi-part Bag-of-Poses. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2013) 479–485
18. Azary, S., Savakis, A.: Grassmannian Sparse Representations and Motion Depth Surfaces for 3D Action Recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2013) 492–499
19. Gao, Z., Song, J.m., Zhang, H., Liu, A.A., Xue, Y.b., Xu, G.p.: Human Action Recognition Via Multi-modality Information. Journal of Electrical Engineering & Technology **9**(2) (2014) 739–748
20. Shen, X., Zhang, H., Gao, Z., Xue, Y., Xu, G.: Human Behavior Recognition Based on Axonometric Projections and PHOG Feature. Journal of Computational Information Systems **10**(8) (2014) 3455–3463
21. Wang, J., Wu, Y.: Learning Maximum Margin Temporal Warping for Action Recognition. In: 2013 IEEE International Conference on Computer Vision (ICCV). (Dec 2013) 2688–2695
22. Yang, X., Tian, Y.: Effective 3D action recognition using EigenJoints. Journal of Visual Communication and Image Representation **25**(1) (2014) 2–11
23. Yuan, J., Wang, J., Liu, Z., Wu, Y.: Learning Actionlet Ensemble for 3D Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(5) (2014) 914–927
24. Chaaraoui, A., Padilla-Lopez, J., Florez-Revuelta, F.: Fusion of Skeletal and Silhouette-Based Features for Human Action Recognition with RGB-D Devices. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW). (Dec 2013) 91–97
25. Chen, C., Liu, K., Kehtarnavaz, N.: Real-time human action recognition based on depth motion maps. Journal of Real-Time Image Processing (2013) 1–9
26. Evangelidis, G., Singh, G., Horaud, R., et al.: Skeletal Quads: Human Action Recognition Using Joint Quadruples. In: Proceedings of the 22nd International Conference on Pattern Recognition (ICPR 2014). (2014)
27. Wang, J., Liu, Z., Wu, Y.: Random Occupancy Patterns. In: Human Action Recognition with Depth Cameras. Springer Briefs in Computer Science. Springer International Publishing (2014) 41–55
28. Luo, J., Wang, W., Qi, H.: Group Sparsity and Geometry Constrained Dictionary Learning for Action Recognition from Depth Maps. In: 2013 IEEE International Conference on Computer Vision (ICCV). (Dec 2013) 1809–1816
29. Zhu, Y., Chen, W., Guo, G.: Fusing Spatiotemporal Features and Joints for 3D Action Recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2013) 486–491
30. Zou, W., Wang, B., Zhang, R.: Human Action Recognition by Mining Discriminative Segment with Novel Skeleton Joint Feature. In: Advances in Multimedia Information Processing PCM 2013. Volume 8294 of Lecture Notes in Computer Science. Springer International Publishing (2013) 517–527
31. Kapsouras, I., Nikolaidis, N.: Action recognition on motion capture data using a dynemes and forward differences representation. Journal of Visual Communication and Image Representation (2014) doi: 10.1016/j.jvcir.2014.04.007

32. Yang, X., Tian, Y.: Super Normal Vector for Activity Recognition Using Depth Sequences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
33. Chen, G., Giuliani, M., Clarke, D., Gaschler, A., Knoll, A.: Action Recognition Using Ensemble Weighted Multi-Instance Learning. In: IEEE International Conference on Robotics and Automation (ICRA). (June 2014)
34. Gowayyed, M.A., Torki, M., Hussein, M.E., El-Saban, M.: Histogram of Oriented Displacements (HOD): Describing Trajectories of Human Joints for Action Recognition. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence. IJCAI'13, AAAI Press (2013) 1351–1357
35. 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 23rd International Joint Conference on Artificial Intelligence. IJCAI'13, AAAI Press (2013) 2466–2472
36. Song, Y., Tang, J., Liu, F., Yan, S.: Body Surface Context: A New Robust Feature for Action Recognition From Depth Videos. IEEE Transactions on Circuits and Systems for Video Technology **24**(6) (2014) 952–964
37. Wang, C., Wang, Y., Yuille, A.: An Approach to Pose-Based Action Recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. (June 2013) 915–922
38. Vieira, A.W., Nascimento, E.R., Oliveira, G.L., Liu, Z., Campos, M.F.: On the improvement of human action recognition from depth map sequences using Space-Time Occupancy Patterns. Pattern Recognition Letters **36**(15) (2014) 221–227
39. Vieira, A., Nascimento, E., Oliveira, G., Liu, Z., Campos, M.: STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Volume 7441 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 252–259
40. Wang, R., Medioni, G., Winstein, C., Blanco, C.: Home Monitoring Musculoskeletal Disorders with a Single 3D Sensor. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2013) 521–528
41. Zhao, X., Li, X., Pang, C., Zhu, X., Sheng, Q.Z.: Online human gesture recognition from motion data streams. In: Proceedings of the 21st ACM International Conference on Multimedia. MM '13, New York, NY, USA, ACM (2013) 23–32
42. Li, X., Sheng, Q., Pang, C., Zhao, X., Wang, S.: Effective approaches in human action recognition. In: 2013 International Conference on Advanced Computer Science and Information Systems (ICACSIS),. (Sept 2013) 1–7
43. Qin, S., Yang, Y., Jiang, Y.: Gesture recognition from depth images using motion and shape features. In: 2nd International Symposium on Instrumentation and Measurement, Sensor Network and Automation (IMSNA). (Dec 2013) 172–175
44. Althloothi, S., Mahoor, M.H., Zhang, X., Voyles, R.M.: Human activity recognition using multi-features and multiple kernel learning. Pattern Recognition **47**(5) (2014) 1800–1812
45. Liang, B., Zheng, L.: Three Dimensional Motion Trail Model for Gesture Recognition. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW). (Dec 2013) 684–691
46. Venkataraman, V., Turaga, P., Lehrer, N., Baran, M., Rikakis, T., Wolf, S.: Attractor-Shape for Dynamical Analysis of Human Movement: Applications in Stroke Rehabilitation and Action Recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2013) 514–520

47. Ellis, C., Masood, S., Tappen, M., LaViola, J., Sukthankar, R.: Exploring the Trade-off Between Accuracy and Observational Latency in Action Recognition. International Journal of Computer Vision **101**(3) (2013) 420–436
48. Wang, J., Liu, Z., Chorowski, J., Chen, Z., Wu, Y.: Robust 3D Action Recognition with Random Occupancy Patterns. In: Proceedings of the 12th European conference on Computer Vision-Volume Part II. Springer Berlin Heidelberg (2012) 872–885
49. Tran, Q., Ly, N.: Sparse spatio-temporal representation of joint shape-motion cues for human action recognition in depth sequences. In: 2013 IEEE International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF). (Nov 2013) 253–258
50. Zanfir, M., Leordeanu, M., Sminchisescu, C.: The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection. In: 2013 IEEE International Conference on Computer Vision (ICCV). (Dec 2013) 2752–2759
51. Tran, Q., Ly, N.: An effective fusion scheme of spatio-temporal features for human action recognition in RGB-D video. In: 2013 International Conference on Control, Automation and Information Sciences (ICCAIS). (Nov 2013) 246–251
52. Rahmani, H., Mahmood, A., Mian, A., Huynh, D.: Real time action recognition using histograms of depth gradients and random decision forests. In: Proceedings of the IEEE Winter Applications of Computer Vision Conference (WACV). (2014)
53. Soh, H., Demiris, Y.: Iterative temporal learning and prediction with the sparse online echo state gaussian process. In: 2012 International Joint Conference on Neural Networks (IJCNN). (June 2012) 1–8
54. Ohn-Bar, E., Trivedi, M.: Joint Angles Similarities and HOG2 for Action Recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2013) 465–470
55. Luo, J., Wang, W., Qi, H.: Spatio-temporal feature extraction and representation for RGB-D human action recognition. Pattern Recognition Letters (2014) doi: 10.1016/j.patrec.2014.03.024
56. Rahmani, H., Mahmood, A., Huynh, D., Mian, A.: Action classification with locality-constrained linear coding. In: Proceedings of the 22nd International Conference on Pattern Recognition (ICPR 2014). (2014)
57. Xia, L., Aggarwal, J.: Spatio-temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),. (June 2013) 2834–2841
58. Climent-Perez, P., Chaaraoui, A.A., Padilla-Lopez, J.R., Florez-Revuelta, F.: Optimal Joint Selection for Skeletal Data from RGB-D Devices Using a Genetic Algorithm. In: Advances in Computational Intelligence. Volume 7630 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 163–174
59. Miranda, L., Vieira, T., Martinez, D., Lewiner, T., Vieira, A., Campos, M.: Real-Time Gesture Recognition from Depth Data through Key Poses Learning and Decision Forests. In: 25th Conference on Graphics, Patterns and Images (SIBGRAPI). (Aug 2012) 268–275
60. Celiktutan, O., Akgul, C.B., Wolf, C., Sankur, B.: Graph-based Analysis of Physical Exercise Actions. In: Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare. MIIRH '13, New York, NY, USA, ACM (2013) 23–32
61. Çeliktutan, O., Wolf, C., Sankur, B., Lombardi, E.: Fast Exact Hyper-graph Matching with Dynamic Programming for Spatio-temporal Data. Journal of Mathematical Imaging and Vision (March 2014) 1–21

62. Chaaraoui, A.A., Padilla-Lepez, J.R., Climent-Perez, P., Florez-Revuelta, F.: Evolutionary joint selection to improve human action recognition with RGB-D devices. Expert Systems with Applications **41**(3) (2014) 786 – 794
63. Azary, S., Savakis, A.: 3D Action Classification Using Sparse Spatio-temporal Feature Representations. In: Advances in Visual Computing. Volume 7432 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 166–175
64. Miranda, L., Vieira, T., Martnez, D., Lewiner, T., Vieira, A.W., Campos, M.F.M.: Online gesture recognition from pose kernel learning and decision forests. Pattern Recognition Letters **39** (2014) 65–73
65. Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., Del Bimbo, A.: Space-Time Pose Representation for 3D Human Action Recognition. In: New Trends in Image Analysis and Processing ICIAP 2013. Volume 8158 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 456–464
66. Barnachon, M., Bouakaz, S., Boufama, B., Guillou, E.: Ongoing human action recognition with motion capture. Pattern Recognition **47**(1) (2014) 238–247
67. Lillo, I., Niebles, J., Soto, A.: Discriminative Hierarchical Modeling of Spatio-Temporally Composable Human Activities. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
68. 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. In: 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2012) 8–13
69. Cottone, P., Re, G., Maida, G., Morana, M.: Motion sensors for activity recognition in an ambient-intelligence scenario. In: 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops). (March 2013) 646–651
70. Chaudhry, R., Ofli, F., Kurillo, G., Bajcsy, R., Vidal, R.: Bio-inspired Dynamic 3D Discriminative Skeletal Features for Human Action Recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2013) 471–478
71. Sabinas, Y., Morales, E., Escalante, H.: A One-Shot DTW-Based Method for Early Gesture Recognition. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Volume 8259 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 439–446