Labeling of Activity Recognition Datasets: Detection of Misbehaving Users

Alessio Vecchio(✉), Giada Anastasi, Davide Coccomini, Stefano Guazzelli, Sara Lotano, and Giuliano Zara

Dip. di Ingegneria dell'Informazione, University of Pisa, Pisa, Italy
alessio.vecchio@unipi.it

Abstract. Automatic recognition of user's activities by means of wearable devices is a key element of many e-health applications, ranging from rehabilitation to monitoring of elderly citizens. Activity recognition methods generally rely on the availability of annotated training sets, where the traces collected using sensors are labelled with the real activity carried out by the user. We propose a method useful to automatically identify misbehaving users, i.e. the users that introduce inaccuracies during the labeling phase. The method is semi-supervised and detects misbehaving users as anomalies with respect to accurate ones. Experimental results show that misbehaving users can be detected with more than 99% accuracy.

Keywords: Activity recognition · Wearable device · Machine learning

1 Introduction

In the last years, we assisted to the proliferation of a large variety of wearable devices such as smart-wristbands, smart-watches, and smart-shoes. All these devices are equipped with sensors and are thus able to provide a rich amount of information about their users. In this context, a significant effort has been devoted to the design and development of methods useful to automatically recognize the activities carried out by people [14,17]. By recognizing the activities of daily living (ADLs), higher-level goals can be achieved. Examples include customization of the environment depending on users' actions (e.g., in a smart-home or in a smart-factory), monitoring of patients' conditions (e.g. to detect an increased sedentary style or falls of elderly citizens) [1,6,7,25], or automated logging of training sessions [4,26]. Many methods rely on machine learning techniques, which must be properly trained to operate successfully. In general, a dataset is collected in a controlled or semi-controlled environment and used to train a system. Then, the trained system is used to recognize the users' activities during the operational phase. Rather obviously, the availability of training datasets characterized by high quality is a necessary condition for obtaining accurate recognition results [27].

G. M. P. O'Hare et al. (Eds.): MobiHealth 2019, LNICST 320, pp. 320–331, 2020.
https://doi.org/10.1007/978-3-030-49289-2_25

Training datasets are generally produced by collecting movement data from a set of users, and then by manually annotating the resulting traces. This process is time consuming and characterized by inaccuracies. The presence of errors in the ground truth negatively impacts the learning phase, and in turn the accuracy of the whole method. Some tools have been proposed to ease the annotation process, e.g. by suggesting the most probable labels to the operator who, most of the time, must simply confirm one of the options [9]. The operator may also be assisted by tools which, during the labeling phase, show a video recorded at the time of the data collection, as an easy way to detect possible errors. Studies demonstrated that assisted labeling is less error-prone and less time-consuming in comparison to a completely manual procedure.

In other cases, the dataset is generated according to a crowdsourcing-based approach, with normal users responsible for both collecting movement data, by means of miniaturized Inertial Measurement Units (IMUs), and labeling the traces. On one hand, crowdsourcing makes possible the creation of large datasets characterized by the presence of a significant number of individuals. On the other hand, the chances of introducing inaccuracies in the dataset get increased by the inclusion of non-professional operators in the process.

In this paper, we propose a method for automatically recognizing the presence of inaccuracies in the labeling phase. In particular, we suppose that users may introduce errors during the labeling phase of their own data. Such inaccuracies can be deliberately introduced by a malicious user who wants to corrupt the dataset, or simply as a consequence of the lack of care during the annotation process. The proposed method relies on one-class classification techniques to understand if one of the users labels his/her data in a way that is significantly different from the other users. Results show that such anomalous users can be identified with more than 99% of accuracy.

2 Related Work

As mentioned, some tools have been proposed in the last years for reducing the effort during the annotation process.

In [16], a data collection tool that allows semi-automated labeling is presented. The tool includes the possibility to manually check and correct labels, and focuses on activity data collected by means of inertial measurements units, pressure insoles, and cameras. The smart annotation tool relies on edge detection, concerning the signal produced by pressure sensors, to achieve a reduction of annotation costs. The tool also helps the operator to synchronize videos and IMU-generated data with the traces produced by pressure sensors. According to the study, the labeling time can be reduced by 83% when using the tool.

The consistency of annotations related to data collected by sensors on a smartphone was studied in [10]. The main goal was to relate the daily behavior of students with their academic performance, using information about their locations and movements. The analyzed data consist of a label, which represents the user's annotation, and the physical location saved by the GPS. First, clusters are obtained by grouping physically close locations. Then, for each user, the consistency of obtained clusters is calculated. Consistency is based on entropy, which considers the number of different labels within a cluster and the number of their occurrences. Considering that the annotations are made by inexperienced users, the results obtained have a reasonable level of consistency (69%). However, by means of semantic analyses, it is possible to obtain a slightly higher level of consistency, equal to 74%. The study mostly focuses on correct labeling of locations.

A method for filtering inaccuracies in a training dataset is described in [2]. In the considered scenario, a trained wearable device – the source device – is used to train a new device – the target device The motivation is that people change wearable devices rather frequently and the knowledge of past devices could be transferred on new ones to reduce the effort required from the user. Initially, source and target device work together while the user carries out his/her activities of daily living. During this period, the predicted label of the source device is transferred to the target device. Then, self-paced learning is used to reduce the impact of inaccuracies [13].

Other tools useful to ease annotation of videos are described in [15,18]. An evaluation of different annotation methods is presented in [23].

In the end, the vast majority of the above mentioned studies, try to reduce the amount of errors introduced during the labeling phase, by assisting the user in different forms. Little attention has been devoted to automatic detection of inaccuracies in datasets, which are used in an always increasing number of studies in the e-health domain.

3 Method

The idea behind the proposed method is to recognize misbehaving, untruthful users as anomalies with respect to a set of truthful ones. In particular, a model of truthful users (TUs) is defined using one-class classification (OCC) methods. Then, the model can be used to recognize untruthful users (UUs) as instances that do not belong to the truthful class.

3.1 One-Class Classification

In machine learning, OCC methods are able to define a model of a single class – the positive class. Training of OCC methods is semi-supervised and requires only samples of the positive class. The absence of non-positive instances during the training phase makes the problem harder with respect to traditional classifiers, as defining the boundaries of the positive class cannot rely on counter-examples [11,12].

OCC methods are particularly useful whenever obtaining non-positive instances is difficult. For example, the normal operational status of an aircraft can be easily observed, while instances of faulty ones are typically unavailable or not common. Another situation where OCC methods are particularly useful is when the negative class is not well-defined: while a news website can be

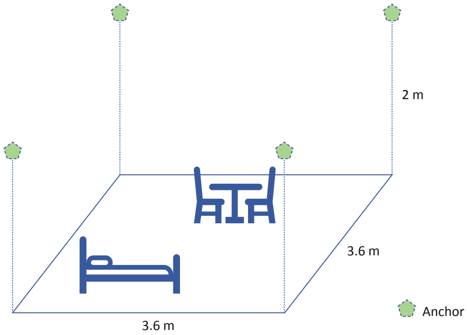

Fig. 1. Room setup.

reasonably identified, all non-news website belong to a such large and diverse set of possible categories that they cannot be easily modeled using traditional classification methods.

The proposed approach relies on OCC mostly because UUs may behave in many different ways, and this makes UUs not easily classifiable. For instance, some malicious users could tag all running activities as walking ones, i.e. they could be systematical in introducing errors during the labeling phase. Sloppy users, on the other hand, could label a given activity as another one, randomly picked, just because of their lack of care.

3.2 Data Collection

We collected a dataset where ten users performed some activities of daily living. Users' movements were captured using both IMUs and Ultra-WideBand (UWB) transceivers. IMUs have been extensively used for this purpose during the last years, as accelerometers and gyroscopes are effective and characterized by reduced costs. UWB transceivers have also been used as they recently became increasingly popular in similar healthcare-related contexts [19,20,22]. In particular, each UWB transceiver is able to determine the distance between itself and another UWB transceiver. If wearable devices are equipped with UWB tranceivers, distance data can be used to obtain information about users' movements.

To collect users' movements we used both Shimmer devices [3], equipped with accelerometers and gyroscopes, and an MDEK Decawave kit [8], whose devices are equipped with transceivers compatible with the IEEE 802.15.4-2011 UWB standard.

In a lab, a room with size 3.6 m × 3.6 m was set up (Fig. 1). Four MDEK sensors were placed at the corners of the room, 2 m above the ground. Such devices operated as "anchors", i.e. nodes whose position is known, and which can be used to compute the position of mobile wearable nodes, called "tags". Each user wore five MDEK devices, and two Shimmer sensors. Devices were attached

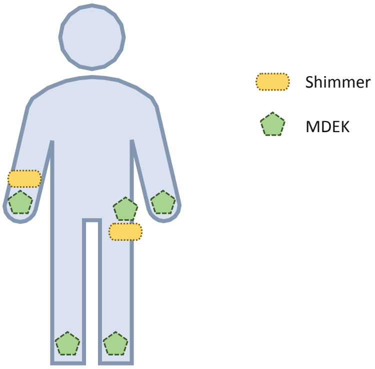

Fig. 2. Position of devices on users' body.

to the users' body according to the scheme illustrated in Fig. 2. MDEK devices were configured to estimate their position with a frequency of 10 Hz, whereas Shimmer devices where set up to collect acceleration and angular velocity at 102.4 Hz. The position of tags in the 3D space of the room was not directly used to understand which was the activity that was currently carried out by the user. 3D positions of tags were used, instead, to compute the distance between couples of wearable devices, e.g. between ankle and wrist or between ankle and pocket. Then, the distances between devices were used to observe user's movements. The rationale for this choice originates from the need to characterize users' movements independently from the position of users in the room. A similar approach was followed in [7], where the reader can find more details.

In the end, for each user, a trace containing the following data was collected: the tri-axial acceleration at the wrist and at the waist, the tri-axial angular velocity at the wrist and at the waist, the ten distances between all the possible couples of UWB-enabled devices (left wrist - left ankle, left wrist - pocket, left wrist - right wrist, etc).

Each user performed six different activities of daily living. Each activity was carried out for one minute. The sequence of activities was: *i*) walking in circle, *ii*) standing in the middle of the room, *iii*) picking up an object repeatedly from the ground, *iv*) sitting, *v*) simulated eating, and *vi*) lying supine.

The main characteristics of the ten users involved in the experiments are shown in Table 1.

3.3 Feature Extraction and Selection

Each user's trace is six minutes long, and contains, as mentioned, 22 signals. Traces have been segmented using fixed size windows, with a duration of 2 s. Then, for each window, a set of functions was computed for all the 22 signals. The adopted functions are: mean, min-max, standard deviation, mean cross ratio, average absolute variation [5], and mean absolute deviation. These functions are

Table 1. Main characteristics of the users involved in the experiments.

User	Height (cm)	Weight (kg)	Age	Gender
1	182	62	29	M
2	158	50	24	F
3	156	65	24	F
4	180	85	23	M
5	182	63	24	M
6	186	78	24	M
7	173	60	28	M
8	176	65	28	M
9	185	62	27	F
10	168	80	24	M

frequently used for signal processing or in the context of activity recognition. Thus for each window, a vector containing 132 features was produced (the feature vector). The number of features was then reduced to 30 using the *relieff* method [21]. This step is generally followed, in activity recognition methods, to avoid overfitting problems and to obtain more efficient systems.

3.4 Identifying Untruthful Users

The resulting dataset contains the feature vectors of all the users. Each feature vector is correctly labelled according to the activity that the user was performing during that time window. The dataset is divided in two parts: one used for training and one used for evaluating the performance of the trained OCC method. In particular, the data of eight users out of ten are used to train an OCC method using only the samples belonging to the positive class, i.e. TUs. The trained OCC method is then evaluated on previously unseen data using the traces of the two remaining users. The OCC method must be evaluated in terms of correct identification of TUs and UUs as truthful and untruthful respectively. To this purpose the data of one of the two remaining users is given as input to the OCC method as it is, and the OCC must identify the user as a truthful one. The data of the last user is transformed to obtain an untruthful one by assigning a wrong, random label to all his/her feature vectors. The transformed data is finally given as input to the trained OCC, which must recognize the user as an untruthful one. This procedure is repeated using all the possible sets of eight users for training, and using all the possible permutations of the remaining two users for the evaluation.

In this context, a true positive means that a TU is classified as a TU, whereas a true negative means that a UU is classified as an UU. Similarly, a false positive means that a UU is classified as a TU, whereas a false negative means that a TU is classified as a UU (Fig. 3).

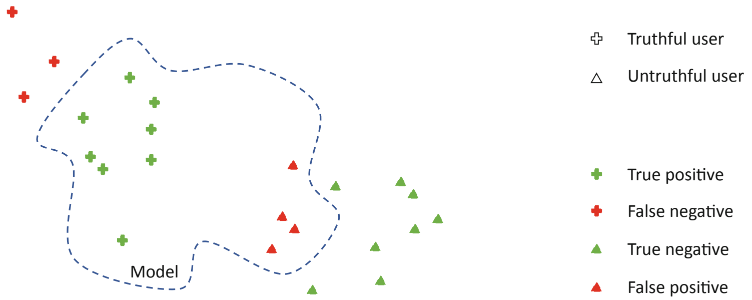

Fig. 3. The model is built using the data of eight users. The remaining data is used to evaluate the trained system.

4 Results

We evaluated the performance of the proposed method when changing some parameters of operation and OCC techniques.

4.1 Impact of the Fraction of Rejected Positive Instances During Training

OCC methods are trained using only positive instances, in our case truthful users. One of the main parameters of OCC methods is the fraction of rejected positive instances during training (*fracrej*). When this parameter is equal to zero, the training phase produces a boundary that includes all positive instances. Such boundary correctly includes all the positive instances provided during the training phase, but it may be prone to generate a number of false positives during the operational phase (some of the positive instances may be particularly far from the "core" of the model). When *fracrej* is greater than zero, a fraction of positive instances are rejected during the training phase. This increases the chances to obtain false negatives during the operational phase, but at the same time reduces the number of false positives (as the boundary is tighter).

We evaluated the performance of the proposed method when *fracrej* is varied in the [0, 0.1] range, when using a Gaussian one-class classifier. Figure 4 shows the obtained false negative rate (FNR) and false positive rate (FPR) of the method. As expected, FPR decreases when *fracrej* increases, whereas FNR increases for larger *fracrej* values. When *fracrej* is equal to zero, the FPR and FNR values are relatively balanced, thus a *fracrej* value equal to zero is used to compute the results presented in Sect. 4.2.

4.2 Combining Results Obtained from Different Windows

The FPR and FNR, obtained by a Gaussian one-class classifier with *fracrej* equal to zero, are 0.20 and 0.14 respectively. Such values suggest that a UU can be reasonably identified, but with some chances to classify a TU as a UU and

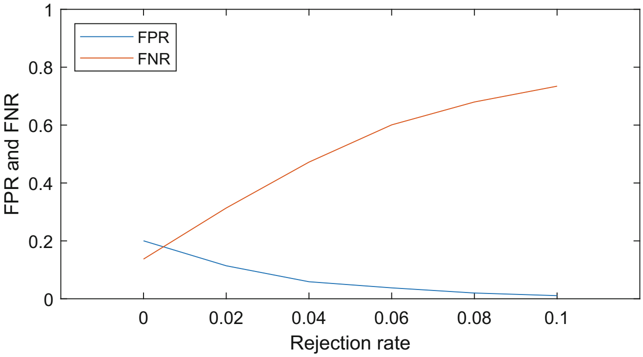

Fig. 4. FPR and FNR when varying the rejection rate.

vice-versa. To improve the performance of the proposed method, we adopted a technique based on majority voting: the results of a set of windows are considered, and the global result is equal to the result obtained in the majority of the windows. Let us define n the number of windows (odd) and k the number of results that indicate the user as an untruthful one. The global result is UU only if $k \geq \lceil \frac{n}{2} \rceil$.

The probability of obtaining k correct results out of n windows can be modelled as a binomial random variable, with probability mass function

$$P(k) = \binom{n}{k} p^k q^{n-k} \tag{1}$$

where p and q are the success and fail probabilities (with $q = 1 - p$). For UUs, q is 0.20 (the FPR on a single window) whereas for TU q is 0.14 (the FNR on a single window). The probability of obtaining the correct result when using n windows is equal to

$$\sum_{k=\lceil n/2 \rceil}^{n} P(k) \tag{2}$$

i.e when the majority of results in the single windows is correct.

Figure 5 shows the results for different values of n, in the [1, 15] interval (obviously, a value of n equal to one corresponds to the case described in Sect. 4.1). When using 15 windows, corresponding to 30 s of user's movements, UUs and TUs can be reliably identified, with a FPR and FNR equal to 0.0042 and 0.0003 respectively.

4.3 Different OCC Techniques

The analysis described in Sect. 4.1 was repeated considering a set of different OCC techniques, besides the Gaussian one. The set of additional methods is: Principal Component Analysis (PCA), Autoencoder, k-means, and Minimum

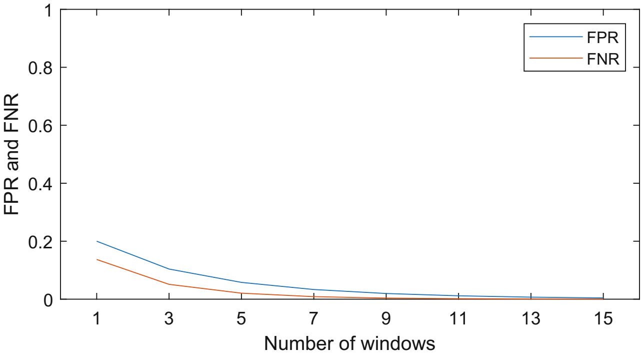

Fig. 5. Incorrect identification (FPR and FNR) of UU and TU with majority voting.

Table 2. Results obtained when using a set of OCC techniques.

OCC method	FPR	FNR	*fracrej*
Gaussian	0.20	0.14	0.0
PCA	0.16	0.15	0.02
Autoencoder	0.19	0.18	0.02
k-means	0.24	0.19	0.08
Minimum Spanning Tree	0.31	0.27	0.08

Spanning Tree [24]. Table 2 shows the obtained results, in terms of FPR and FNR. For each OCC technique, also the *fracrej* value that provided the best result is indicated. The overall best result is achieved by the OCC version of PCA, with FPR and FNR values equal to 0.16 and 0.15 respectively. When the majority voting technique is applied to the OCC PCA classifier, the final values of FPR and FNR are equal/below $1 \cdot 10^{-3}$ (when using 15 windows). This confirms that correct identification of TUs and UUs is possible with high accuracy when using just 30 s of data.

5 Conclusion

Automatic recognition of user's activities by means of wearable devices is a key element of many e-health applications, ranging from rehabilitation to monitoring of elderly citizens. Human activity recognition generally relies on supervised machine learning, where an annotated dataset is used to train the system. An annotated dataset requires the users or the operators to manually specify a label associated to the activity performed during a specific time interval of the training traces.

The proposed method is able to reliably identify untruthful users (or operators), i.e. the ones who associate wrong labels to trace segments. Given a set of positive examples, the method is able to detect untruthful users as anomalies, thus without the need of counter-examples. As far as we know, this problem received very little attention despite the importance of training datasets, used as ground truth, in the context of human activity recognition.

Presented results have been obtained under the assumption that a set of truthful users is initially available to train the OCC method. Future work will study the impact caused by the presence of a fraction of untruthful users also in the initial set. Finally, to better assess the performance of the proposed method, further studies will include a larger dataset, both in terms of users and duration of performed activities.

Acknowledgment. This work was partially funded by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence).

References

1. Abbate, S., Avvenuti, M., Bonatesta, F., Cola, G., Corsini, P., Vecchio, A.: A smartphone-based fall detection system. Perv. Mobile Comput. **8**(6), 883–899 (2012). https://doi.org/10.1016/j.pmcj.2012.08.003, http://www.sciencedirect.com/science/article/pii/S1574119212000983, special Issue on Pervasive Healthcare
2. Bao, Y., Chen, W.: Automatic model construction for activity recognition using wearable devices. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 806–811, March 2018. https://doi.org/10.1109/PERCOMW.2018.8480411
3. Burns, A., et al.: Shimmer^{TM} - a wireless sensor platform for noninvasive biomedical research. IEEE Sens. J. **10**(9), 1527–1534 (2010). https://doi.org/10.1109/JSEN.2010.2045498
4. Chambers, R., Gabbett, T.J., Cole, M.H., Beard, A.: The use of wearable microsensors to quantify sport-specific movements. Sports Med. **45**(7), 1065–1081 (2015). https://doi.org/10.1007/s40279-015-0332-9
5. Cola, G., Avvenuti, M., Vecchio, A., Yang, G., Lo, B.: An unsupervised approach for gait-based authentication. In: 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN), pp. 1–6, June 2015. https://doi.org/10.1109/BSN.2015.7299423
6. Cola, G., Vecchio, A., Avvenuti, M.: Improving the performance of fall detection systems through walk recognition. J. Ambient Intell. Humanized Comput. **5**(6), 843–855 (2014). https://doi.org/10.1007/s12652-014-0235-x
7. Aliperti, A., et al.: Using an indoor localization system for activity recognition. In: Sugimoto, C., Farhadi, H., Hämäläinen, M. (eds.) BODYNETS 2018. EICC, pp. 233–243. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-29897-5_19
8. Decawave: www.decawave.com. Accessed 15 July 2019
9. Diete, A., Sztyler, T., Stuckenschmidt, H.: A smart data annotation tool for multi-sensor activity recognition. In: 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 111–116, March 2017. https://doi.org/10.1109/PERCOMW.2017.7917542

10. Giunchiglia, F., Zeni, M., Bignotti, E., Zhang, W.: Assessing annotation consistency in the wild. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 561–566, March 2018. https://doi.org/10.1109/PERCOMW.2018.8480236
11. Khan, S.S., Madden, M.G.: One-class classification: taxonomy of study and review of techniques. Knowl. Eng. Rev. **29**(3), 345–374 (2014)
12. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: Proceedings of The Twenty-First International Conference on Machine Learning, p. 62. ACM (2004)
13. Kumar, M.P., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.) Advances in Neural Information Processing Systems, vol. 23, pp. 1189–1197. Curran Associates, Inc. (2010). http://papers.nips.cc/paper/3923-self-paced-learning-for-latent-variable-models.pdf
14. Lara, O.D., Labrador, M.A.: A survey on human activity recognition using wearable sensors. IEEE Commun. Surv. Tutorials **15**(3), 1192–1209 (2013). https://doi.org/10.1109/SURV.2012.110112.00192
15. Liu, C., Freeman, W.T., Adelson, E.H., Weiss, Y.: Human-assisted motion annotation. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, June 2008. https://doi.org/10.1109/CVPR.2008.4587845
16. Martindale, C.F., Roth, N., Hannink, J., Sprager, S., Eskofier, B.M.: Smart annotation tool for multi-sensor gait-based daily activity data. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 549–554, March 2018. https://doi.org/10.1109/PERCOMW.2018.8480193
17. Mukhopadhyay, S.C.: Wearable sensors for human activity monitoring: a review. IEEE Sens. J. **15**(3), 1321–1330 (2015). https://doi.org/10.1109/JSEN.2014.2370945
18. Palotai, Z., et al.: LabelMovie: semi-supervised machine annotation tool with quality assurance and crowd-sourcing options for videos. In: 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 1–4, June 2014. https://doi.org/10.1109/CBMI.2014.6849850
19. Qi, Y., Soh, C.B., Gunawan, E., Low, K.S., Maskooki, A.: A novel approach to joint flexion/extension angles measurement based on wearable UWB radios. IEEE J. Biomed. Health Inform. **18**(1), 300–308 (2013)
20. Qi, Y., Soh, C.B., Gunawan, E., Low, K.S., Maskooki, A.: Using wearable UWB radios to measure foot clearance during walking. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5199–5202. IEEE (2013)
21. Robnik-Šikonja, M., Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. Mach. Learn. **53**(1), 23–69 (2003). https://doi.org/10.1023/A:1025667309714
22. Shaban, H.A., El-Nasr, M.A., Buehrer, R.M.: Toward a highly accurate ambulatory system for clinical gait analysis via UWB radios. IEEE Trans. Inf. Technol. Biomed. **14**(2), 284–291 (2010). https://doi.org/10.1109/TITB.2009.2037619
23. Szewcyzk, S., Dwan, K., Minor, B., Swedlove, B., Cook, D.: Annotating smart environment sensor data for activity learning. Technol. Health Care **17**(3), 161–169 (2009)
24. Tax, D.: Ddtools, the data description toolbox for Matlab, January 2018. Version 2.1.3

25. Vecchio, A., Cola, G.: Fall detection using ultra-wideband positioning. In: 2016 IEEE SENSORS, pp. 1–3, October 2016. https://doi.org/10.1109/ICSENS.2016.7808527
26. Vecchio, A., Mulas, F., Cola, G.: Posture recognition using the interdistances between wearable devices. IEEE Sens. Lett. **1**(4), 1–4 (2017). https://doi.org/10.1109/LSENS.2017.2726759
27. Yordanova, K., Krüger, F., Kirste, T.: Providing semantic annotation for the CMU grand challenge dataset. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 579–584, March 2018. https://doi.org/10.1109/PERCOMW.2018.8480380