Multi-Layer Cross Loss Model for Zero-Shot Human Activity Recognition

Tong Wu[1,2], Yiqiang Chen[1,2,3](✉), Yang Gu[1,2,3], Jiwei Wang[1,2], Siyu Zhang[1,2,3], and Zhanghu Zhechen[1,2]

[1] The Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
{wutong17s,yqchen,guyang,wangjiwei}@ict.ac.cn, astro6974@gmail.com, zhangzhanghuzhechen@outlook.com
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Peng Cheng Laboratory, Shenzhen 518055, China

Abstract. Most existing methods of human activity recognition are based on supervised learning. These methods can only recognize classes which appear in the training dataset, but are out of work when the classes are not in the training dataset. Zero-shot learning aims at solving this problem. In this paper, we propose a novel model termed **M**ulti-**L**ayer **C**ross **L**oss **M**odel (MLCLM). Our model has two novel ideas: (1) In the model, we design a multi-nonlinear layers model to project features to semantic space for that the deeper the network is, the better the network can fit the data's distribution. (2) A novel objective function combining mean square loss and cross entropy loss is designed for the zero-shot learning task. We have conduct sufficient experiments to evaluate the proposed model on three benchmark datasets. Experiments show that our model outperforms other state-of-the-art methods significantly in zero-shot human activity recognition.

Keywords: Human activity recognition · Zero-shot learning · Cross loss

1 Introduction

Human Activity Recognition (HAR) has appealed much attention in recent years because of its usage in many applications, such as fall detection [1], game consoles [2], etc. HAR mainly depends on 2 kinds of signals: video camera and inertial sensors integrated in wearable devices. With the developments of wearable devices, inertial sensors have been widely employed in the field of HAR. The main reasons are threefold: (1) The wearable devices with inertial sensors are convenient while the camera is usually fixed in a specific place. (2) The inertial sensors' data requires less storage while the camera usually needs large memory

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12084, pp. 210–221, 2020.
https://doi.org/10.1007/978-3-030-47426-3_17

for visual data. Besides the storage, processing the videos is also costly. (3) The inertial sensors just record the user's information, while the video camera contains information of others in the same place, so inertial sensors have inherent advantage of protecting user's privacy and are more target-specific.

Due to the fast developments in deep neural networks, the performance of human activity recognition has enhanced significantly in recent years. But most of the HAR methods are supervised learning methods [3–5]. They can just recognize the classes appeared in training dataset but are incapable of recognizing the classes not appeared in training dataset. Nevertheless, the training dataset can't contain all the activities, because on the one hand every individual can do plenty of activities, and we can't collect all the activities before training stage, on the other hand, it's extremely expensive to annotate an activity and label the training data. Recognizing the classes not appeared in training dataset is defined as zero-shot learning problem which is first proposed in [6]. This study provided a formal framework to solve the problem and a zero-shot learning example for the activity decoding task. In the setting of zero-shot learning, the classes in the training dataset (seen classes) and classes in the testing dataset (unseen classes) are disjoint. So in order to recognize the unseen classes, it needs extra information about the seen and unseen classes. The extra information can be defined as semantic space. After the model have captured the relations between the feature space and semantic space of the seen classes, it can transfer the relations to unseen classes. The main idea of zero-shot learning is to correlate the unseen classes with the seen classes via the semantic space. The semantic space can be divided into 2 categories. One is the text vector space, which includes the word-embedding of the classes' names and the text description of these classes. The other is the attribute space, where the attributes are defined by human beings.

In recent years, there are many zero-shot learning methods proposed in human activity recognition field. However, these methods have defects in different aspects. In [7], it used the Support Vector Machine (SVM) classifier as base classifier to detect attributes, and each attribute needed an SVM classifier. Once the attribute space became larger, the model would become extremely complex. In [8], the model needed to use the testing dataset during the training stage, so the model could just be used in a fixed number of unseen classes. Once a new instance which belonged to neither seen classes nor unseen classes appeared, it would be out of work. In this paper, we present a novel model for human activity recognition in the zero-shot learning task. It outperforms other state-of-the-art methods and is termed as **M**ulti-**L**ayer **C**ross **L**oss **M**odel (MLCLM). In this model, it learns a multi-fully connected layers model to project features to the attribute space, and the instance is predicted through a similarity classifier (SC). What's more, our model is not sensitive to the number of unseen classes, for that it just learns the projection between the feature space and semantic space, and once the semantic representations of new classes are defined, our model can deal with these new classes. The major contributions of this paper are as follows:

– To the best of our knowledge, we are the first to introduce the word-embedding into zero-shot human activity recognition.

- We propose a novel model and split the problem into two sub-problems, which can optimize the model in two spaces.
- Sufficient experiments on three benchmark datasets show that the proposed model outperforms other state-of-the-art methods and through these results, we analyze how the attribute and word-embedding impact the performance of our model.

The rest of the paper is organized as follows: In Sect. 2, we review related work of the zero-shot learning and HAR. In Sect. 3, the proposed model is explained in detail. In Sect. 4, we evaluate the performance of our model. In Sect. 5, we summarize our work and discuss future work.

2 Related Works

Supervised human activity recognition has achieved great success in recent years. Many researches have been completed in this area [9–11]. Cao et al. [9] presented an integrated framework that used non-sequential machine learning tools to achieve high performance on activity recognition with multi-modal data. Wang et al. [10] constructed a decision tree to classify different walking patterns based on relations between gait phases. Shi et al. [11] proposed a dynamic coordinate transformation approach to recognize activity with valid recognition results. Most of these proposed methods based on supervised learning. They can only recognize the classes in the training dataset. An instance that does not belong to any classes in the training dataset will not be recognized. So these methods are limited to a fixed number of classes.

Zero-shot learning aims at figuring out the defects of supervised learning methods. In zero-shot learning methods, the methods are mainly divided into 2 categories. One category is inductive zero-shot learning [12–15], where the model has no information about the unseen classes except the semantic information. In other words, the unseen classes are uncertain. Kodirov et al. [12] proposed a semantic autoencoder model with a novel objective function to reconstruct the features after the projection from features to semantic space. Romeara et al. [13] applied the mean square loss and Frobenius norm as the objective function to learn a bilinear compatibility model. Liu et al. [14] employed the temperature calibration in the prediction probability and introduced an additional entropy loss to maximize the confidence of classifying the seen data to unseen classes. Another category is transductive zero-shot learning [16–18], where it can use the information of the unseen classes, including feature space and semantic space. The above methods are in the field of computer vision. In zero-shot learning for human activity recognition scenario, several methods have been proposed [7,15,19]. Chen et al. [7] proposed an inductive zero-shot learning method. In [7], for each attribute, it learned an attribute probability classifier (SVM). It then got the class representation in the semantic space and predicted label with the maximum a posteriori estimate (MAP). Cheng et al. [19] proposed an extended work following [7], it changed the SVM to CRF to improve the performance.

Wang et al. [15] proposed a model which learned a nonlinear compatibility function between the semantic space and feature space and classified an instance to the class with the highest score.

3 Proposed Method

In this section, we will explain the proposed zero-shot learning model for human activity recognition explicitly. We call it **M**ulti-**L**ayer **C**ross **L**oss **M**odel (MLCLM). The input of our model is the features extracted from the inertial sensors' data.

3.1 Problem Definition

Unlike supervised human activity recognition methods, the problem of zero-shot learning is defined as follows: the training dataset $D_{train} = (x_i, y_i)_{i=1}^{N}$ contains N labeled training instances from the seen classes $S = \{S_1, S_2, S_3, ..., S_m\}$. The unseen classes are denoted by $U = \{U_1, U_2, U_3, ..., U_k\}$, whose instances are not in the training dataset. Seen classes and unseen classes are disjoint, $S \cap U = \emptyset$. Each class $c \in S \cup U$ owns a semantic representation, denoted by $a_c \in R^A$, where the dimension of semantic space is A.

3.2 Data Preprocessing

The data received by sensors can't be used directly in our model. It should be preprocessed first. To deal with the raw data from sensor readings, we adopt the sliding window mechanism which is commonly used in time series data to segment the data, as shown in Fig. 1. Each window of the time series data is defined as an instance.

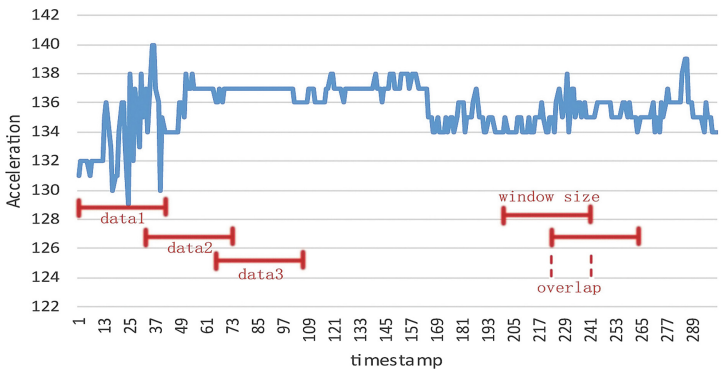

Fig. 1. Data segmentation

Then we collect statistical features from the segments, and the features are usually the mean value, standard deviation and time, etc. As the time series data in the window size may belong to different labels (for example, when the subject is changing activities), we discard these instances when labeling the instances after segmentation.

3.3 Proposed Model

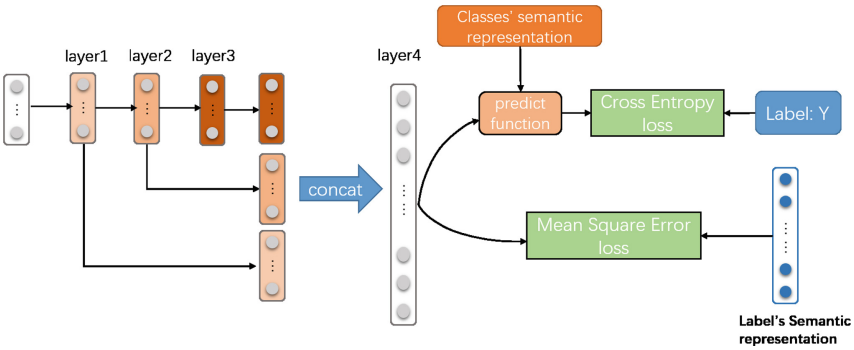

Fig. 2. The structure of the proposed MLCLM

In our model, we define four fully connected layers to project the features extracted from the sensor readings to the semantic space. As shown in Fig. 2. The model can be formulated as Eq. 1.

$$h_i = \begin{cases} \phi_1(f_1(x_j)) & i = 1 \\ \phi_i(f_i(h_{i-1})) & i = 2, 3 \\ \phi_4(f_4([h_1, h_2, h_3]) & i = 4 \end{cases} \tag{1}$$

The input of the first fully connected layer is the input of the model, and the input of the second and the third fully connected layers is the output of the previous layer. The concatenation of the former three layers is the input of the forth fully connected layer. We believe that former fully connected layers' output can also contribute to classification as we conduct a traditional classification experiment on the datasets. we find that even we use the input of the layer1 to do a traditional classification task, we can still get good results.

Prediction function: The classification is predicted by the similarity classifier (SC). After the multi-fully connected layers' model, the input features x_i are projected to the semantic space, defined as $\phi(x_i)$. Then the similarities with all the classes T ($T = S$ during the training stage, while $T = U$ during the testing stage) are calculated, as shown in the equation (2).

$$SC(x_i, c_p) = \frac{\phi(x_i) \cdot a^p}{\sum_j^A a_j^p} \tag{2}$$

$c_p \in T$ and a^p is the semantic representation of the class c_p (a^p is a semantic vector), which can be obtained in many ways such as defined by human beings or pretrained word-embedding. The numerator is a dot product of the two vectors. After calculating the similarity with all the classes, we choose the most similar class as the prediction of input x_i:

$$\hat{y}(x_i) = argmax_{c_p} SC(x_i, c_p) \tag{3}$$

Cross loss: In our model, we consider the zero-shot learning problem as a regression problem and a classification problem. The regression problem occurs after projecting the input features to the semantic space. In this regression problem, the mean square loss is chosen as an objective function, defined as M:

$$M = \sum_{i=1}^{N} \sum_{j}^{A} (\phi(x_i)_j - a_j^i)^2 \tag{4}$$

$\phi(x_i)$ is the output of the multi-fully connected layers model and a^i is the semantic representation of x_i's true label. The classification problem is occurred after the prediction function. We choose the cross entropy loss as the objective function in this problem, for that the cross entropy loss is an effective solution for multi-class classification. Before we apply the cross entropy, we first transform the similarity to probability as:

$$p(c_p|x_i) = \frac{exp(SC(x_i, c_p)}{\sum_{c_j} exp(SC(x_i, c_j))} \tag{5}$$

where $c_p, c_j \in T$. Then we apply the cross entropy:

$$L = -\sum_{i=1}^{N} \sum_{c_p} y_{i,c_p} * log(p(c_p|x_i)) \tag{6}$$

So the optimization of the zero-shot learning problem can be formulated as:

$$\min_{\phi} M + L + \gamma\Omega(f) \tag{7}$$

where the $\Omega(f)$ is the regularization of the multi-fully connected layers model. In our code, we implemented the Ω regularization by weight decay.

4 Experiment

In order to evaluate the proposed model, we perform extensive experiments on three benchmark datasets. And we compare our model with other state-of-the-art methods. In this section, we will introduce the three benchmark datasets and experiments.

4.1 Datasets and Preprocessing

The three benchmark datasets are TU Darmstadt dataset (TUD) [20], Physical Activity Monitoring Data Set (PAMAP2) [21], and Opportunity Activity Recognition Data Set (OPP) [22]. When segmenting the datasets, we follow the sliding window strategy proposed in their papers. In TUD, we use a sliding window of 30 s with 15 s overlap between two adjacent windows. After the segmentation, we extract mean value and standard deviation as features in each dimension from the sensor readings in a window. Besides the 3 axies acceleration data from the sensors placed in the subjects, we add the time as another feature to the feature vector. In PAMAP2, we segment the dataset using a sliding window of 5.12 s with 1 s overlap between two adjacent windows. After the segmentation, we extract mean value, standard deviation as features in each dimension from the sensor readings in a window. In OPP, we follow the sliding window mechanism adopted by [23], where the window size is 1 s, and the overlap is 0.5 s. After the segmentation, we extract the mean and standard deviation as features from each dimension.

In the TUD, we discard the 'unlabeled' data, and in PAMAP2, we discard the data with label '0' (transient activities), while in the OPP dataset, we discard the data of 'drill's file.

To adopt zero-shot learning to HAR, the semantic space is essential for the three datasets. As presented in Sect. 2, there are two categories of semantic space: attribute space and text vector space, so we conduct experiments on the two categories.

4.2 Experimental Results

The evaluation metric in the experiments is the average accuracy of each class, defined as *average_acc_per_class* in Eq. (8) [24], which is commonly used in zero-shot learning. This is due to the fact that the activities are unbalanced in the dataset, e.g. Fig. 3 shows the amounts of classes in the PAMAP2. The amount of 'computer work' activity is extremely greater than in other activities. So if the model predicts all the testing instances as this class, the average accuracy on all the instances will be fine but the model has no robustness.

$$average\_acc\_per\_class = \frac{1}{k} * \sum_{i}^{k} \frac{N^i_{correct}}{N^i_{total}} \tag{8}$$

where the $N^i_{correct}$ indicates the number of correct predictions of the class i, the N^i_{total} indicates the total instances of the class i, and the k indicates the number of unseen classes.

1) **Word-embedding experiments:** As presented in Sect. 2, there are two categories of semantic space: attribute space and text vector space. The attribute space is defined manually by experts with domain knowledge, which is costly. And once the activity classes change, it needs extra efforts to redefine the attributes for the new activity. So in this experiment, we firstly introduce the pretrained word2vec's [25] word-embedding on part of Google News

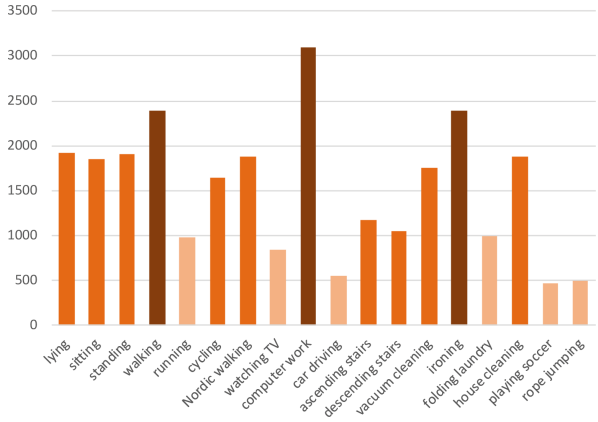

Fig. 3. The amounts of instances in each class of PAMAP2 dataset

dataset as the semantic space in the HAR field. In the word-embedding, every word has a 300-dimensional vector. But in the TUD and PAMAP2, each activity's name may contain several words, so we calculate the mean value of all words in the activity name as the word-embedding of the activity $w^{mean} = \{w_1^{mean}, w_2^{mean}, w_3^{mean}, ..., w_{300}^{mean}\}$ for the activity:

$$w_j^{mean} = \frac{\sum_i^I w_j^i}{I}, j = 1, 2, 3, ..., 300 \tag{9}$$

w_j indicates the jth word's word-embedding; I indicates the amount of the words in the activity name.

Table 1. Results of word-embedding experiments

Methods	TUD	PAMAP2	OPP
DCN [14]	27.42%	32.97%	46.03%
ESZSL [13]	20.57%	27.66%	**52.37%**
SAE [12]	17.71%	31.79%	44.55%
NCBM [15]	20.15%	27.49%	30.80%
MLCLM	**29.89%**	**54.93%**	**51.50%**

We compare the proposed model with the methods proposed in [12–15]. Wang et al. proposed a zero-shot learning method called **NCBM** [15] for HAR. Several representative methods of zero-shot learning in computer vision are also chosen: **DCN** [14], **ESZSL** [13], **SAE** [12] and we transfer them into HAR. In the experiments, we choose the class splitting strategy of the train and test datasets proposed in [15]. Here we adopt the 5-fold cross validation strategy to evaluate

the performance of the methods and our model. Here we didn't choose the **DAP** [26] method, because the **DAP** used the SVM classifier as the attribute classifier, but the word-embedding is a continuous number, so the SVM classifier is not befitting.

The results are shown in Table 1. From the results, we can see that our method(**MLCLM**) outperforms other state-of-the-art methods significantly in the TUD and PAMAP2, while in the OPP, we are very closed to the best.

2) Attribute experiment: Besides using the word-embedding as the semantic space, we also conduct experiments with attributes defined by human beings. The results are shown in Table 2. From the results, we can see that our model (**MLCLM**) outperforms others over more than 23%–32% in PAMAP2 and is the best in OPP. While in the TUD, our method is also closed to the best in the table.

Table 2. Results of comparison experiments

Methods	TUD	PAMAP2	OPP
DAP [26]	16.08%	29.24%	30.93%
DCN [14]	24.69%	37.85%	47.52%
ESZSL [13]	27.36%	32.58%	60.55%
SAE [12]	34.73%	34.46%	61.41%
NCBM [15]	31.07%	38.06%	28.44%
MLCLM	**31.87%**	**61.28%**	**62.30%**

We analyze that the reason why the **MLCLM** outperforms the others is that in the **MLCLM**, the optimization has two constraints: On the one hand, the mean square loss can minimize the gap between the conversion of the features and the semantic space. On the other hand, the cross entropy loss can further optimize the results on classification. So combining with the two losses, our model can be optimized both on semantic space and the classification results.

In the **DCN**, there are two optimization functions in the **DCN**: the cross entropy loss of predicting seen data on seen classes and the entropy of predicting seen data on unseen classes. However, the second entropy loss can cause misclassification of seen data on unseen classes, which may cause an underfitting problem. The **SAE** uses an autoencoder model, and it reconstructs the features which are projected to the semantic space. However, even when the reconstruction is perfect, the classification results don't benefit from the reconstruction. In the **DAP**, it assigns an SVM classifier for each attribute. Assembling all results of SVM classifiers, the features are projected to the semantic space, and the prediction is made by the maximum a posteriori estimate (MAP). However, each SVM classifier only concentrates on its attribute and ignores the relations of the attributes and the truth label. In other words, it doesn't optimize directly on the classification results, which can cause overfitting on the attribute classification but underfitting on the class classification results. The **ESZSL** also

has the problem that it is only optimized on the semantic space, but not on the class classification results. The **NCBM** learns a nonlinear compatibility function, which gives the compatibility scores of the input and all class prototypes, and the class with the highest score is the result of the input. The optimization of **NCBM** is the hinge loss, which is not sensitive to the outliers, however, the unseen classes are outliers to the seen classes, so this loss can not employ the outliers' information.

3) Comparison experiments: To evaluate how the attribute and word-embedding impact the performance of our **MLCLM**, we compare the performance when using them as semantic space, and Fig. 4. shows results. From the results, we can see that the performance of attribute outperforms the word-embedding. To find out why the attribute outperforms the word-embedding, we take the PAMAP2 dataset as a representative: we calculate the Pearson Product-Moment Correlation Coefficient (PPMCC) of these classes' semantic representations in attributes and word2vec. We find out that the PPMCC of attributes is larger than the PPMCC of word2vec. According to zero-shot learning, the model needs to learn the relations between the feature space and the semantic space in the seen classes, and then it transfers the relations to the unseen classes. So the more relevant between the seen classes and unseen classes in the semantic space, the better the model performs in the testing stage. This conclusion can guide us on how to define the attribute space better for the zero-shot learning task.

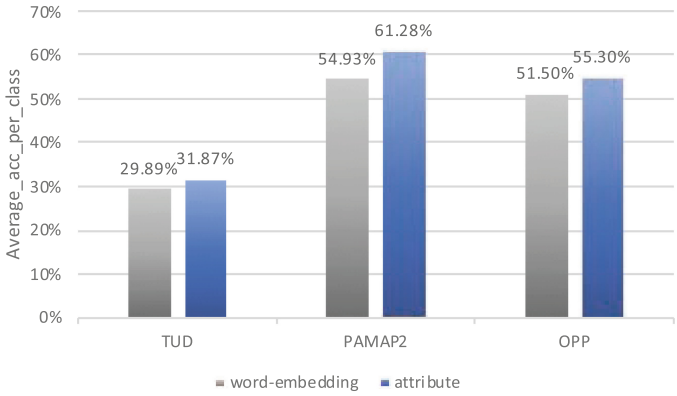

Fig. 4. Results of the word-embedding and attribute as semantic space

5 Conclusion and Future Work

In this paper, we propose a **M**ulti-**L**ayer **C**ross **L**oss **M**odel for zero-shot learning in human activity recognition. Sufficient experiments validate that the proposed model is effective with both attributes and word-embedding as semantic space. In the future, we will upgrade our model for better performance in zero-shot

learning, and apply our method to other fields like computer vision and natural language processing. As the results shown in Fig. 4, the results are not very ideal when using word-embedding as semantic space, so in the future, we will explore more researches on the employment of word-embedding as it needs fewer efforts of human beings.

Acknowledgement. This work is supported by National Key Research & Development Program of China No.2017YFC0803401; Natural Science Foundation of China No. 61902377; Beijing Natural Science Foundation No. 4194091; R & D Plan in Key Field of Guangdong Province No.2019B010109001; Alibaba Group through Alibaba Innovative Research (AIR) Program.

References

1. Khojasteh, S., Villar, J., Chira, C., González, V., de la Cal, E.: Improving fall detection using an on-wrist wearable accelerometer. Sensors **18**(5), 1350 (2018)
2. Direkoğlu, C., O'Connor, N.E.: Temporal segmentation and recognition of team activities in sports. Mach. Vis. Appl. **29**(5), 891–913 (2018). https://doi.org/10.1007/s00138-018-0944-9
3. Inoue, M., Inoue, S., Nishida, T.: Deep recurrent neural network for mobile human activity recognition with high throughput. Artif. Life Rob. **23**(2), 173–185 (2017). https://doi.org/10.1007/s10015-017-0422-x
4. Cao, L., Wang, Y., Zhang, B., Jin, Q., Vasilakos, A.V.: GCHAR: an efficient group-based context-aware human activity recognition on smartphone. J. Parallel Distrib. Comput. **118**, 67–80 (2018)
5. Asghari, P., Soelimani, E., Nazerfard, E.: Online human activity recognition employing hierarchical hidden markov models. arXiv preprint arXiv:1903.04820 (2019)
6. Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M.: Zero-shot learning with semantic output codes. In: Advances in neural information processing systems, pp. 1410–1418 (2009)
7. Cheng, H.T., Sun, F.T., Griss, M., Davis, P., Li, J., You, D.: Nuactiv: recognizing unseen new activities using semantic attribute-based learning. In: Proceeding of the 11th annual international conference on Mobile systems, applications, and services, pp. 361–374. ACM (2013)
8. Zheng, V.W., Hu, D.H., Yang, Q.: Cross-domain activity recognition. In: Proceedings of the 11th international conference on Ubiquitous computing, pp. 61–70. ACM (2009)
9. Cao, H., Nguyen, M.N., Phua, C., Krishnaswamy, S., Li, X.: An integrated framework for human activity classification. In: UbiComp, pp. 331–340 (2012)
10. Wang, J.S., Lin, C.W., Yang, Y.T.C., Ho, Y.J.: Walking pattern classification and walking distance estimation algorithms using gait phase information. IEEE Trans. Biomed. Eng. **59**(10), 2884–2892 (2012)
11. Shi, D., Wu, Y., Mo, X., Wang, R., Wei, J.: Activity recognition based on the dynamic coordinate transformation of inertial sensor data. In: 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), pp. 1–8. IEEE (2016)

12. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3174–3183 (2017)
13. Romera-Paredes, B., Torr, P.: An embarrassingly simple approach to zero-shot learning. In: International Conference on Machine Learning, pp. 2152–2161 (2015)
14. Liu, S., Long, M., Wang, J., Jordan., M.I.: Generalized zero-shot learning with deep calibration network. In: Advances in Neural Information Processing Systems, pp. 2005–2015 (2018)
15. Wang, W., Miao, C., Hao, S.: Zero-shot human activity recognition via nonlinear compatibility based method. In: Proceedings of the International Conference on Web Intelligence, pp. 322–330. ACM (2017)
16. Rohrbach, M., Ebert, S., Schiele, B.: Transfer learning in a transductive setting. In: Advances in neural information processing systems, pp. 46–54 (2013)
17. Kodirov, E., Xiang, T., Fu, Z., Gong, S.: Unsupervised domain adaptation for zero-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2452–2460 (2015)
18. Fu, Y., Sigal, L.: Semi-supervised vocabulary-informed learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5337–5346 (2016)
19. Cheng, H.T., Griss, M., Davis, P., Li, J., You, D.: Towards zero-shot learning for human activity recognition using semantic attribute sequence model. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 355–358. ACM (2013)
20. Huynh, T., Fritz, M., Schiele, B.: Discovery of activity patterns using topic models. UbiComp **8**, 10–19 (2008)
21. Reiss, A., Stricker, D.: Introducing a new benchmarked dataset for activity monitoring. In: 2012 16th International Symposium on Wearable Computers, pp. 108–109. IEEE (2012)
22. Roggen, D., et al.: Collecting complex activity datasets in highly rich networked sensor environments. In: 2010 Seventh international conference on networked sensing systems (INSS), pp. 233–240. IEEE (2010)
23. Hammerla, N.Y., Halloran, S., Plötz, T.: Deep, convolutional, and recurrent models for human activity recognition using wearables. arXiv preprint arXiv:1604.08880 (2016)
24. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 2251–2265 (2018)
25. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
26. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE Trans. Pattern Anal. Mach. Intell. **36**(3), 453–465 (2013)