Unfamiliar Dynamic Hand Gestures Recognition Based on Zero-Shot Learning

Jinting Wu[1,2(✉)], Kang Li[1,2], Xiaoguang Zhao[1,2], and Min Tan[1,2]

[1] The State Key Laboratory of Management and Control for Complex System, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
{wujinting2016,likang2014,xiaoguang.zhao,min.tan}@ia.ac.cn

Abstract. Most existing robots can recognize trained hand gestures to interpret user's intent, while untrained dynamic hand gestures are hard to be understood correctly. This paper presents a dynamic hand gesture recognition approach based on Zero-Shot Learning (ZSL), which can recognize untrained hand gestures and predict user's intention. To this end, we utilize a Bidirectional Long-Short-Term Memory (BLSTM) network to extract hand gesture feature from skeletal joint data collected by Leap Motion Controller (LMC). Specifically, this data is used to construct a novel dynamic hand gesture dataset for human-robot interaction application. Twenty common hand gestures are included and fifteen concrete semantic attributes are condensed. Based on these features and semantic attributes, a Semantic Autoencoder (SAE) is employed to learn a mapping from feature space to semantic space. By matching the most similar semantic information, the unfamiliar hand gestures are recognized as correct as possible. Experimental results on our dataset indicate that the proposed approach can effectively identify unfamiliar hand gestures.

Keywords: Dynamic hand gesture recognition
Bidirectional Long-Short-Term Memory (BLSTM)
Zero-Shot Learning (ZSL) · Semantic Autoencoder (SAE)
Leap Motion Controller (LMC)

1 Introduction

Recently, hand gesture recognition has been widely applied in various fields, such as medical technology [14,24], sign language recognition [3,8], virtual reality and human-robot interaction [2,9,17]. Specially, in human-robot interaction, hand gesture, as one of the most intuitive and efficient interactive interfaces, can help a person with speech barrier communicate with the robot [17], remote control the robot [9] and express intention [2]. Therefore, a robot with the capability of recognizing hand gestures becomes more practical and valuable.

However, a limitation of existing hand gesture recognition algorithms is that they need to learn gestures from large amounts of labeled image data. Thus, they

L. Cheng et al. (Eds.): ICONIP 2018, LNCS 11305, pp. 244–254, 2018.
https://doi.org/10.1007/978-3-030-04221-9_22

can only classify familiar hand gestures based on training dataset. In real human-robot interaction, the robot may encounter some unfamiliar hand gestures. Under these circumstances, the robot needs to have the ability to guess what meanings the unseen hand gestures convey. Fortunately, ZSL methods provide a solution for identifying unseen categories.

ZSL relies on a labeled training set of seen classes and the semantic relationship between the seen and unseen classes. Seen and unseen classes are usually related in a semantic embedding space. The semantic relationships between classes can be measured by a distance in this space. In the recognition task, the class label of a test sample is assigned to the nearest unseen class prototype in the semantic space. Currently, ZSL is mainly applied in 2D image recognition for object classification [5,10,11,23]. The studies about hand gesture recognition [18] are very rare.

In this paper, we present a novel unfamiliar dynamic hand gesture recognition method based on ZSL. First, hand and finger skeletal joint data is collected by a LMC and used to build a novel hand gesture dataset involving twenty common hand gestures and fifteen concrete semantic attributes. Then, we utilize a BLSTM network [7] to extract hand gesture features and employ SAE [10] to analyze the semantic information. By matching the predicted semantic representation and the semantic prototypes, the unfamiliar gestures can be inferred. Finally, experimental results on the novel hand gesture dataset demonstrate effectiveness of the proposed method.

The remainder of the paper is organized as follows. Related work on hand gesture recognition is reviewed in Sect. 2. Afterwards, the unfamiliar hand gesture recognition approach is elaborated in Sect. 3. Section 4 provides the experimental results and analysis. Finally, conclusions and future work are summarized in Sect. 5.

2 Related Work

Traditional work on hand gesture recognition mostly uses information captured by data gloves [4] or 2D digital cameras [22]. However, the data gloves are not user-friendly, and the 2D images include limited information for dynamic hand gesture recognition. In recent years, depth sensors, such as the Leap Motion controller (LMC) [6,16,19] and Microsoft Kinect [12,20,25], have widely used in hand gesture recognition because they can contribute rich 3D information to enhance the accuracy. Especially, LMC is cheaper and more portable, and has higher localization precision (which is about 0.2 mm [26]). Abundant 3D hand data, such as palm positions, hand directions and skeletal joint positions, can be easily collected by a LMC without extra computational work. For instance, Lu et al. used palm direction, palm normal and fingertip data captured from a LMC to recognize dynamic hand gestures, and reached an accuracy of 95.0% for the Handicraft-Gesture dataset [16]. Chen et al. utilized a LMC to acquire the motion trajectory of 36 hand gestures, and the accuracy of SVM approach is 98.24% [6].

In addition to determining the source of the data, the recognition method is also important for hand gesture recognition. Wang et al. utilized a Hidden Markov Model (HMM) to estimate motion trajectory of hand gesture in a service robot system [9]. Lu et al. first recognized dynamic hand gestures using Hidden Conditional Random Field (HCRF) which was only applied in speech recognition [16]. Tang et al. employed Deep Neural Networks (DNNs) to extract robust features and precisely recognize hand postures [25]. All of the aforementioned hand gesture recognition methods have a common drawback that they cannot identify unfamiliar hand gestures. How to do it? ZSL algorithms make it possible.

Since Lampert et al. first proposed the attribute-based classification approach, which was used to identify new classes based on attribute representation [11], a large amount of ZSL models have been proposed one after another to improve the performance of unseen class recognition. Paredes et al. proposed a ZSL approach which adopted two linear layers to construct relationships between features, attributes and categories [23]. However, that method left a large of dimensions of the semantic space unconstrained. To solve this problem, Morgado et al. combined two main strategies of ZSL, which are Recognition using independent semantics (RIS) and Recognition using semantic embeddings (RULE) [21]. However, the algorithm is limited to its model complexity and computational cost. In our proposed method, we use a linear SAE for ZSL [10], which achieved state-of-the-art performance and had lower computational cost.

Most ZSL methods in object recognition application extract features by Convolutional Neural Networks (CNNs). However, CNNs are not suitable for extracting dynamic spatio-temporal sequential hand gesture features. Considering that recurrent neural networks (RNNs) can encoder temporal information of dynamic hand gesture sequences [15], we use RNNs to pre-process our original skeletal joint data. In practice, Long Short-term Memory (LSTM), as a special RNN architecture which replaces traditional artificial neurons in the hidden layer with memory cells [15], can overcome the issue of gradient vanishing and error blowing up. The Bidirectional LSTM network involves two hidden LSTM layers (forwards and backwards) to store and process both past and future information [7]. Thus, we use a BLSTM network to extract features of hand data.

3 Unfamiliar Hand Gesture Recognition Approach Based on Zero-Shot Learning

3.1 Overview of Unfamiliar Hand Gesture Recognition Approach

The brief flow of our approach is shown in Fig. 1. Three modules are included: data collection, feature extraction and ZSL. First, hand gesture data are captured by LMC and pre-processed. Then, a BLSTM network is employed to extract hand gesture features from the pre-processed data. Finally, we utilize a SAE model for ZSL to learn a mapping from the feature space to the semantic space. By comparing the distances between the estimated semantic representations and the prototypes in the semantic space, hand gestures can be recognized.

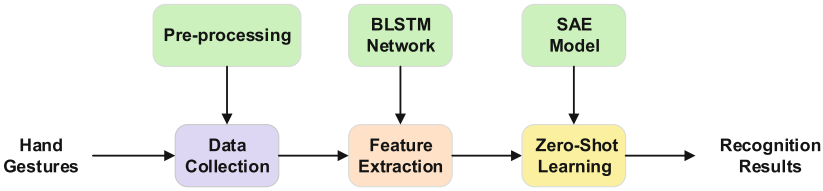

Fig. 1. Brief flow of the approach architecture.

3.2 Collecting Hand Gesture Data

We collect hand gesture data by a LMC. As shown in Fig. 2, the data frames of LMC involve much information, such as palm positions, skeletal joint positions, and so on. We choose the following information on a single right hand as the input of our recognition system:

1. Palm center position in 3D space.
2. The pitch, yaw and roll of the hand, which are calculated from the hand direction vector and palm normal vector.
3. The 3D positions of finger skeletal joints.

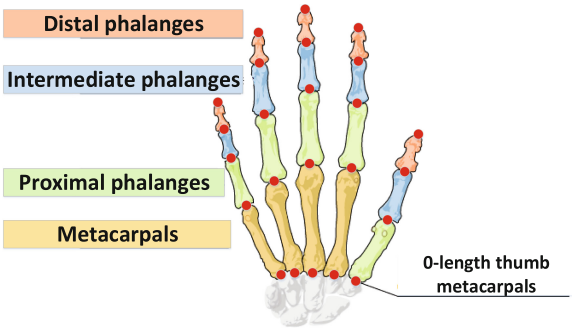

Fig. 2. Hand bones captured by the LMC. The red circles are finger skeletal joints. (Color figure online)

We record hand gesture data with 50 Hz sampling rate. Pre-processing mainly includes three steps. First, we eliminate the invalid data and select a fixed number of frames from each sequence. Then, to decrease the influence of different hand location, the skeletal joint positions are replaced by the positions in relation to the palm center position. Finally, these data are normalized to the interval [0, 1] based on z-score.

3.3 Feature Extraction

To better analyze the time correlation among sequential frame, a BLSTM network is used to extract spatio-temporal features from hand gesture data. The structure of this network is shown in Fig. 3. It is comprised of one input layer, one BLSTM layer and one output layer. The BLSTM layer concludes two LSTM layers (a forward one and a backward one), which can respectively deal with the past and future spatio-temporal context [13]. The BLSTM layer is fully connected to the input layer, and the outputs of the BLSTM layer are high-level feature expressions. The size of the output layer is equal to the number of labeled hand gestures. We use the softmax classifier to predict recognition results.

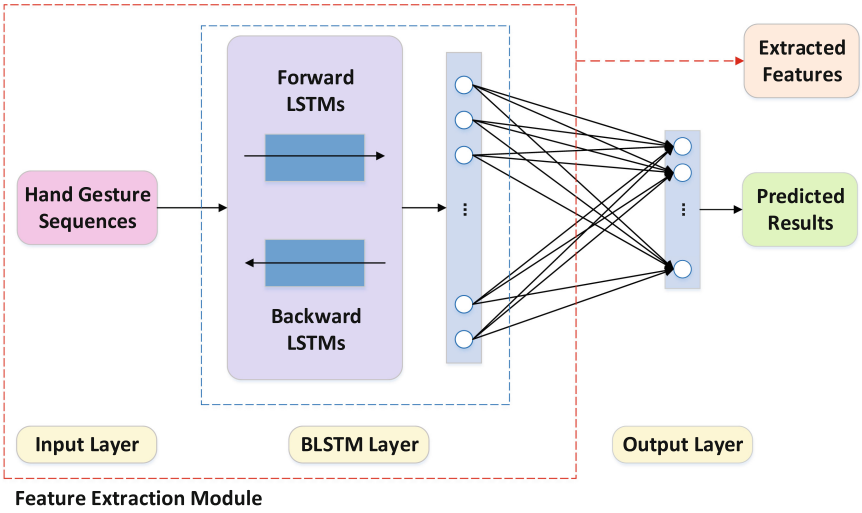

Fig. 3. The structure of the BLSTM network. (Color figure online)

In the training stage, we adopt the five-fold cross validation to select the best model. The labeled data is divided into five parts. Each part is chosen as the validation set without repetition. Meanwhile, the other four parts constitute the training set to train the network. In the feature extraction stage, we delete the output layer of the trained BLSTM model and put the labeled and unlabeled data into this network to extract the feature vectors (see the output of red dashed box in Fig. 3).

3.4 ZSL for Unfamiliar Hand Gesture Recognition

To recognize unfamiliar hand gesture, we need to learn high-level semantic representations from extracted feature vectors in Sect. 3.3. Our method is based on the SAE for ZSL [10], which achieved state-of-the-art performance currently. We employ the simplest antoencoder which is linear and only has one hidden

layer. An antoencoder contains an encoder and a decoder. The encoder projects features into the hidden layer which represents the attribute space in our experiment, and the decoder projects the attribute vectors back to the feature space to reconstruct the original features.

During training, the input hand gesture features are denoted as $\mathbf{X}_Y = \{\mathbf{x}_i\} \in \mathbb{R}^{d\times N}$ and the semantic attributes are denoted as $\mathbf{S}_Y = \{\mathbf{s}_i\} \in \mathbb{R}^{k\times N}$, where $\mathbf{x}_i$ is a d-dimensional feature vector extracted from the i-th training sample, and $\mathbf{s}_i$ is a k-dimensional corresponding semantic attribute vector of the i-th training sample. The goal of ZSL algorithm is to obtain a projection matrix $\mathbf{W} \in \mathbb{R}^{k\times d}$ which can describe the mapping from the feature space to the semantic attribute space. The objective function is formulated as:

$$\min_{\mathbf{W}} \left\|\mathbf{X} - \mathbf{W}^T\mathbf{S}\right\|_F^2 + \lambda \left\|\mathbf{W}\mathbf{X} - \mathbf{S}\right\|_F^2 \tag{1}$$

This unconstrained optimization problem can be transformed into solving a Sylvester equation by using Bartels-Stewart algorithm [1].

During testing, the test hand gesture features $\mathbf{X}_Z = \{\mathbf{x}_i\}$ and the attribute prototypes of unseen classes $\mathbf{S}_Z$ are provided to predict the labels of untrained samples. Based on encoder projection matrix $\mathbf{W}$ obtained in training, we can project a new test sample $\mathbf{x}_i \in \mathbf{X}_Z$ to the semantic space by $\hat{\mathbf{s}}_i = \mathbf{W}\mathbf{x}_i$. Then, we predict an ideal hand gesture class label with minimum distance between estimated semantic representation $\hat{\mathbf{s}}_i$ and the projection prototypes $\mathbf{S}_Z$:

$$\Phi\left(\mathbf{x}_i\right) = \arg\min_j D\left(\hat{\mathbf{s}}_i, \mathbf{S}_{Z_j}\right) \tag{2}$$

where $\mathbf{S}_{Z_j}$ is the attribute vector of the j-th unseen class, D is the L2 distance function, and $\Phi\left(\cdot\right)$ returns the class label of the sample. More theoretical derivation and proof details about SAE model can be found in [10].

4 Experimental Results and Analysis

4.1 Experimental Setting

Dataset. Because of the lack of open 3D dynamic hand gesture data captured by the LMC, we build a novel dataset for our recognition task. Sixteen training hand gestures and four test hand gestures contained in our dataset are shown in Fig. 4. For each hand gesture class, we collected 50 data sequences, and each sequence consists of 50 frames. Particularly, a frame contains hand direction, palm center and 25 skeletal joint positions. Therefore, each frame is described as an 81-dimensional vector.

Parameter Settings. For training the BLSTM network, the number of training epochs is set to 100, and both batch size and the number of forward and backward LSTM neurons are set to 64. We select the cross-entropy function as the loss function, and the Adam optimization algorithm is utilized to minimize the loss. After training, we extract a 128-dimensional feature vector from each sample before the output layer. The output features of training set and test set are included in our dataset for the subsequent ZSL.

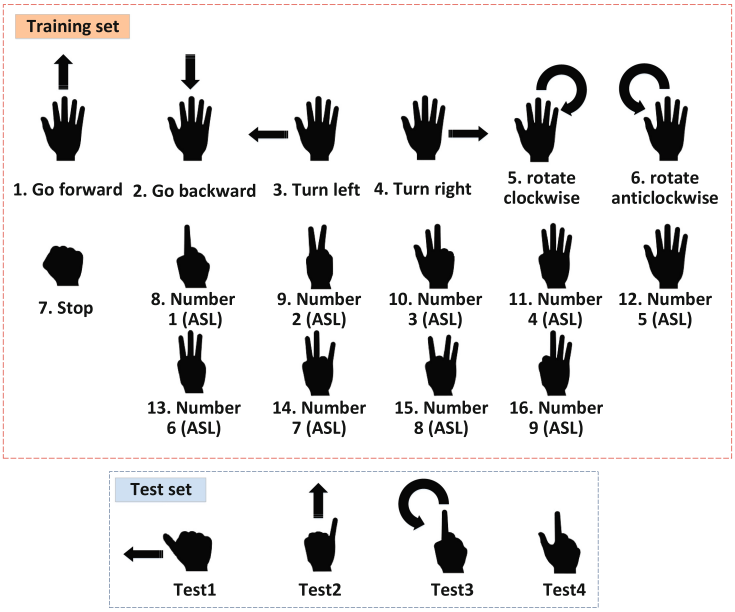

Fig. 4. Hand gestures in our dataset.

Semantic Representation. We condense fifteen semantic attributes about various hand gestures. The correspondence between hand gesture classes and semantic attributes is shown in Table 1.

Table 1. The semantic attribute description

Lable	Hand gesture name	Dynamic state of hand		Shape of trajectory		Direction of trajectory						Finger bending				
		Rest	Motion	Straight line	Circle	Clockwise	Anticlockwise	Forward	Backward	Left	Right	Thumb	Index finger	middle finger	Ring finger	Pinky finger
0	go forward	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0
1	go backward	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0
2	turn left	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0
3	turn right	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0
4	rotate clockwise	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
5	rotate anticlockwise	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0
6	stop	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1
7	number 1	1	0	0	0	0	0	0	0	0	0	1	0	1	1	1
8	number 2	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1
9	number 3	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1
10	number 4	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
11	number 5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	number 6	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1
13	number 7	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0
14	number 8	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0
15	number 9	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
17	test1	0	1	1	0	0	0	0	0	1	0	0	1	1	1	1
18	test2	0	1	1	0	0	0	1	0	0	0	1	1	1	1	0
19	test3	0	1	0	1	0	1	0	0	0	0	1	0	1	1	1
20	test4	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1

4.2 Experimental Results and Analysis

In this section, we conducted extensive evaluation on unfamiliar hand gesture recognition task on our dataset. In the first experiment, we conducted comparisons with two state-of-the-art ZSL models, which are the Embarrassingly Simple Zero-Shot Learning (ESZSL) [23] and the Synthesized Classifiers (SYNC) [5], respectively. The qualitative evaluation results are shown in Fig. 5, and the confusion matrices are shown in Fig. 6. We can observe that our model can significantly outperform other methods. We also evaluate the computational cost of these three ZSL methods. Table 2 shows that for model training and testing, our method is the fastest.

In the second experiment, the effect of various size of training dataset is evaluated. We randomly delete fixed number classes from original sixteen classes in

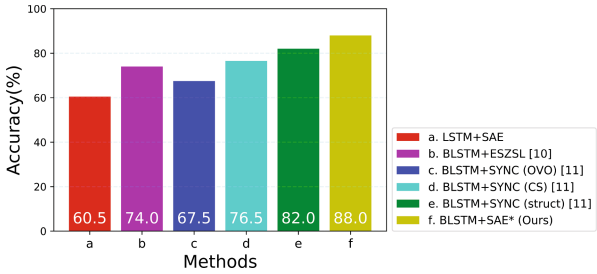

Fig. 5. The qualitative evaluation results of different recognition methods.

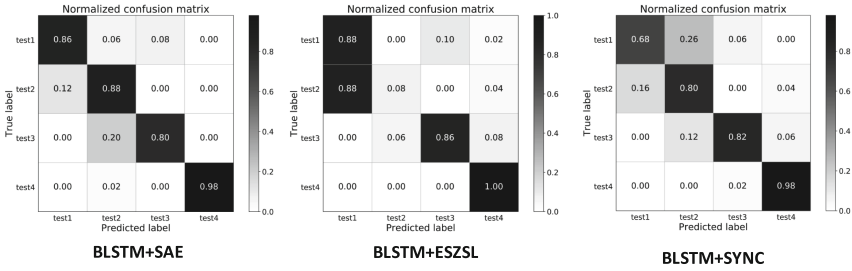

Fig. 6. The confusion matrices of different ZSL models.

Table 2. Comparative evaluation on computation cost

Method	Training time (s)	Test time (s)
ESZSL	1.1311	0.147546
SYNC (CS)	1.4941	0.43829
SYNC (OVO)	1.8176	0.41969
SYNC (struct)	1.0786	0.48054
SAE	0.437094	0.018566

training dataset, and calculate the average accuracies of 10 repeated experiments. Average accuracies with different numbers of deleted classes are shown in Fig. 7. From the trend of curve, we can see that the average accuracies will decrease when deleting more classes from the training dataset. Because more attributes cannot be learned, the more unlabeled classes will be not recognized. In despite of deleting 5 training classes, we achieve the recognition accuracy of 37.5% , which still demonstrates the effectiveness of our method.

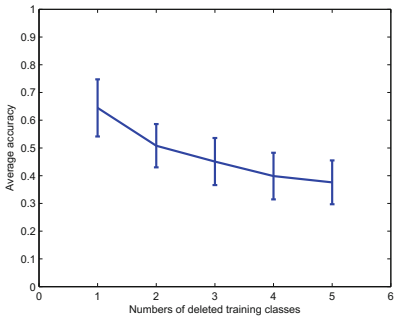

Fig. 7. Average accuracy of different numbers of training classes. Vertical bars indicate ±1 standard deviation.

4.3 Discussion

The experimental results have indicated that the proposed method can well recognize the unfamiliar hand gestures. However, there are some improving space. In the process of attribute design, more attributes can be added to describe more complicated hand gestures. Deeper BLSTM network can be utilized to extract more sophisticated hand gesture features.

5 Conclusion and Future Work

In this paper, we present a novel unfamiliar hand gesture recognition approach based on ZSL. We collect hand and finger joint data from the LMC and construct a hand gesture dataset with semantic information. The BLSTM network is used to extract features and the SAE model for ZSL is built to infer semantic description of unlabeled hand gestures. By matching the predicting semantic information and ground truth, the unfamiliar hand gestures can be inferred. Finally, the experimental results verify that our method achieves state-of-the-art performance and has lower computational cost than other methods.

In the future, the proposed method will be applied to the real-life human-robot interaction system. We plan to realize a real-time online hand gesture recognition interface which can make the robot correctly understand user's intention even though they use unfamiliar hand gestures.

Acknowledgments. This work is partially supported by the National Natural Science Foundation of China under Grants 61673378 and 61421004.

References

1. Bartels, R.H., Stewart, G.W.: Solution of the matrix equation AX+ Xb = C [F4]. Commun. ACM **15**(9), 820–826 (1972)
2. Van den Bergh, M., et al.: Real-time 3D hand gesture interaction with a robot for understanding directions from humans. In: RO-MAN, pp. 357–362. IEEE (2011)
3. Bheda, V., Radpour, D.: Using deep convolutional networks for gesture recognition in American sign language. arXiv preprint arXiv:1710.06836 (2017)
4. Camastra, F., De Felice, D.: LVQ-based hand gesture recognition using a data glove. In: Apolloni, B., Bassis, S., Esposito, A., Morabito, F. (eds.) Neural Nets and Surroundings. Smart Innovation, Systems and Technologies, vol. 19, pp. 159–168. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-35467-0_17
5. Changpinyo, S., Chao, W.L., Gong, B., Sha, F.: Synthesized classifiers for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5327–5336 (2016)
6. Chen, Y., Ding, Z., Chen, Y.L., Wu, X.: Rapid recognition of dynamic hand gestures using leap motion. In: 2015 IEEE International Conference on Information and Automation, pp. 1419–1424. IEEE (2015)
7. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
8. Ji, Y., Liu, C., Gong, S., Cheng, W.: 3D hand gesture coding for sign language learning. In: 2016 International Conference on Virtual Reality and Visualization (ICVRV), pp. 407–410. IEEE (2016)
9. Ke, W., Li, W., Ruifeng, L., Lijun, Z.: Real-time hand gesture recognition for service robot. In: 2010 International Conference on Intelligent Computation Technology and Automation (ICICTA), vol. 2, pp. 976–979. IEEE (2010)
10. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. arXiv preprint arXiv:1704.08345 (2017)
11. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 951–958. IEEE (2009)
12. Lefebvre, G., Berlemont, S., Mamalet, F., Garcia, C.: BLSTM-RNN based 3D gesture classification. In: Mladenov, V., Koprinkova-Hristova, P., Palm, G., Villa, A.E.P., Appollini, B., Kasabov, N. (eds.) ICANN 2013. LNCS, vol. 8131, pp. 381–388. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40728-4_48
13. Lefebvre, G., Berlemont, S., Mamalet, F., Garcia, C.: Inertial gesture recognition with BLSTM-RNN. In: Koprinkova-Hristova, P., Mladenov, V., Kasabov, N.K. (eds.) Artificial Neural Networks. SSB, vol. 4, pp. 393–410. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-09903-3_19
14. Li, W.J., Hsieh, C.Y., Lin, L.F., Chu, W.C.: Hand gesture recognition for post-stroke rehabilitation using leap motion. In: 2017 International Conference on Applied System Innovation (ICASI), pp. 386–388. IEEE (2017)
15. Lipton, Z.C., Berkowitz, J., Elkan, C.: A critical review of recurrent neural networks for sequence learning. arXiv preprint arXiv:1506.00019 (2015)
16. Lu, W., Tong, Z., Chu, J.: Dynamic hand gesture recognition with leap motion controller. IEEE Signal Process. Lett. **23**(9), 1188–1192 (2016)

17. Luo, R.C., Wu, Y.C.: Hand gesture recognition for human-robot interaction for service robot. In: 2012 IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), pp. 318–323. IEEE (2012)
18. Madapana, N., Wachs, J.P.: A semantical & analytical approach for zero shot gesture learning. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 796–801. IEEE (2017)
19. Marin, G., Dominio, F., Zanuttigh, P.: Hand gesture recognition with leap motion and kinect devices. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 1565–1569. IEEE (2014)
20. Molchanov, P., Gupta, S., Kim, K., Kautz, J.: Hand gesture recognition with 3D convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–7 (2015)
21. Morgado, P., Vasconcelos, N.: Semantically consistent regularization for zero-shot recognition. In: CVPR, vol. 9, p. 10 (2017)
22. Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: a survey. Artif. Intell. Rev. **43**(1), 1–54 (2015)
23. Romera-Paredes, B., Torr, P.: An embarrassingly simple approach to zero-shot learning. In: International Conference on Machine Learning, pp. 2152–2161 (2015)
24. Shen, J., Luo, Y., Wang, X., Wu, Z., Zhou, M.: GPU-based realtime hand gesture interaction and rendering for volume datasets using leap motion. In: 2014 International Conference on Cyberworlds (CW), pp. 85–92. IEEE (2014)
25. Tang, A., Lu, K., Wang, Y., Huang, J., Li, H.: A real-time hand posture recognition system using deep neural networks. ACM Trans. Intell. Syst. Technol. (TIST) **6**(2), 21 (2015)
26. Weichert, F., Bachmann, D., Rudak, B., Fisseler, D.: Analysis of the accuracy and robustness of the leap motion controller. Sensors **13**(5), 6380–6393 (2013)