Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Discriminative Recognition of Point Cloud Gesture Classes through One-Shot Learning*

Joshua Owoyemi
ZMP Co, Inc.
Koishikawa, Bunkyo, Tokyo
joshua.owoyemi@zmp.co.jp

Naoya Chiba and Koichi Hashimoto
Graduate School of Information Sciences, Tohoku University
Sendai, Miyagi, Japan
chiba@ic.is.tohoku.ac.jp, koichi@tohoku.ac.jp

***Abstract*— In this paper, we introduce a one-shot learning approach to extend learned point cloud gesture categories into recognizing newly introduced categories after training the original model. This approach is based on learning the discrimination between gesture classes of point cloud data, making it possible to recognize new classes without retraining the original deep neural network (DNN) model. We develop a temporal variant of the PointNet model referred to as Temporal PointNet or TPoinNet, which consumes a sequence of raw point clouds from a low-cost depth sensor and outputs the class of the sequence. We use a multitask strategy where the model learns the classification of a point cloud sequence and a Euclidean space discrimination between the classified sequence and another sequence of a different class. The new model is able to classify and map the point clouds sequence inputs into a Euclidean space where the distances between the gesture sequences correspond to the gesture similarities. We present results on a point cloud dataset and the MSR Action 3D dataset showing the discrimination of new gesture categories with a high precision.**

***Index Terms*— Point Cloud, One-shot learning, Gesture Recognition**

I. Introduction

Gesture recognition is a very important area in Human Computer Interaction (HCI), with potential in facilitating interaction between humans and artificial intelligent systems such as robots, virtual reality, augmented reality and smart devices. In this paper, we propose a strategy that can make it possible to learn new classes of gestures using few examples. This approach is called One-shot learning. The significance of this approach is that users interacting with a system can easily teach a new gesture class without retraining the gesture recognition model. One-shot learning is learning a class from a single labeled example [1]. That is, given a test sample we will make a prediction after observing a single example of each possible class of the test instance. There is also zero-shot [2] and k-shot learning [3], where k signifies the number of examples needed to learn a new class.

Many attempts have been made to extend learning from a DNN model to new classes or problem domain. An earlier idea is called transfer learning [4], where learned parameters of an already trained model is used to initialize or act as bottleneck layers in the new model. This works well but the new model has to be retrained to adapt the learned weights into the new problem. This method only saves time on retraining a model from scratch but does not solve the problem of the need for a huge amount of data also for retraining. Since the introduction of one-shot learning in [5], there has been more notable examples such as Siamese Neural Networks applied on hand written digits [6], Memory-Augmented Neural Networks for sequential data [7], prototypical networks proposed in [8] to learn classification by calculating distance to the prototype expression of each class, the graph neural network [9] that extends existing neural network methods for processing the data represented in graph domains, the neural turing machines [10] that can learn strategies to store expressions in memory and use those expressions for prediction, and Matching Networks for image class recognition [1], which learns a network that maps a small labeled support set and an unlabeled example to its label.

As far as we know, the one-shot learning method has mainly been applied to image data, usually for recognition of classes from few examples. In this paper we aim to apply one-shot learning to discriminating new categories of point cloud gesture data.

One-shot gesture learning aims at learning gesture classes with one or few samples. Most research in this area in past have focused on hand engineered features to maximize feature extraction from limited depth data samples. In [11], one-shot gesture learning was applied on RGBD data by building gray pyramid, depth pyramid and optical flow pyramids as scale for each gray frame and depth frame. Then interest regions are extracted according to the variance of the optical flow and variance is calculated in horizontal and vertical direction. A different approach in [12] was to compute a space time descriptor using the standard deviation of the depth images of a gesture as well as the motion history image. Similar to aforementioned is the extraction of spatiotemporal features from RGBD data using mixed features around sparse keypoints (MFSK) [13].

The most similar reviewed work to our approach is [14],

* This work was supported by JSPS KAKENHI Grant Numbers JP16H06536, JP18J20111 and ACT-I, JST Grant Numbers JPMJPR16UH.

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

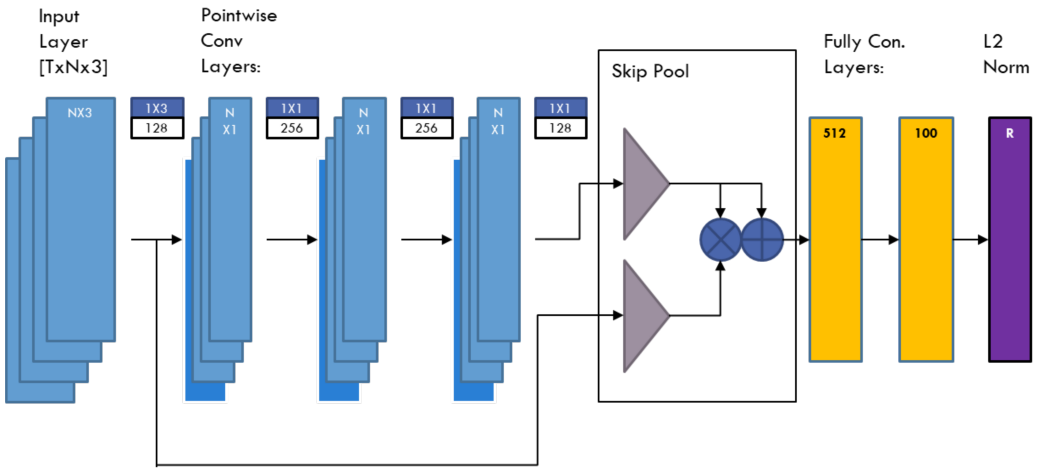

Fig. 1: Our developed TPointNet. The inputs are streams of point cloud data. ReLu activations are used and dropout layers are used after the fully connected layers.

where a DNN gesture recognition model is trained with the intention of using it as a transferable model to extract features from original classes of gestures. Then a discriminative recognition is carried out with Euclidean distance measure between root features and test features, while also updating the representation of the root features as the model is exposed to more examples. However, our approach is more end-to-end as our model does both feature extraction and class discrimination in one step of the pipeline. Our optimization strategy actually helps to eliminate the hassle of finding a suitable discrimination threshold as this is part of the objective function, so our model learns to separate different classes by the enforced distance in the objective function.

II. Gestures Class Embedding with Temporal PointNet

The works in [15] [16], where a case was made for using point cloud data, came up with an approach to recognize gestures from point cloud sequences. In this paper we improve on the approach by developing the Temporal PointNet (TPoinNet) model, a very simple model, based on the original PointNet model [17] to be more efficient and accurate for point cloud gesture classification. The TPoinNet as shown in Figure 1 consumes raw point cloud sequence to output the corresponding gesture class. It consist of 4 convolutional layers followed by 3 fully connected layers. For max pooling, which represents a symmetric function on the points, we added pooling from the first layer to pooling after the fourth layer (skip pool). This is aimed at allowing the model to pool global features as well as local features from the input. We collected a dataset of point cloud hand gestures using a table top depth sensor. The results from the classification from TPoinNet model is given in Table I. More on the dataset is given in a later section II-B.

Our one shot learning approach is achieved by comparing the Euclidean distance $D_e = E(x_i, \hat{x})$ between the $\mathbb{R}^d$ embedding of gesture samples and then optimizing the TPointNet model $f(x_i)$, using the triplet loss function [18] to minimize the distance between similar gesture and maximize the distance between dissimilar gesture samples.

The triplet loss function is expressed as:

$$Loss_{triplet} = \sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+ \tag{1}$$

Where a signifies the anchor or test sample, p a positive sample which is in the same class with a, and n a negative sample which is not in the same class as a. α is a margin that is enforced between positive and negative pairs. We want to ensure that a gesture x_i^a (anchor) of a specific class is closer to all other gestures x_i^p (positive) of the same gesture than it is to any gesture x_i^n (negative) of any other gesture. On the last layer of the model is an L_2 normalization layer which aims at producing a normalized output representing the features of the samples in the Euclidean space. The network is trained such that the squared L_2 distances in the embedding space directly correspond to sample similarity.

TABLE I: Number of Points Comparison

Number of Input Points	**Classification Accuracy** (%)
512	93.41
1024	94.0
2048	94.67

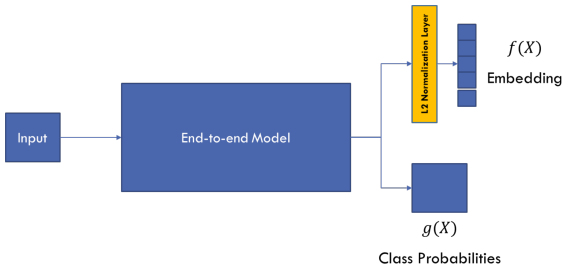

Fig. 2: Our multitask model for simultaneous classification and new class recognition. The model outputs are class probabilities and Euclidean space embedding representation for the input sample. The embedding representation is compared with known categories in order to find out if it belongs to a known category or a new category.

A. Class Embedding Comparisons using Euclidean Distance Threshold

In order to recognize a new class, we will compare the similarity of new samples with known classes k by using the Euclidean distance $D_e = E(x_i, \hat{x}) = \|f(x_k) - f(\hat{x}_k)\|$ between the features of known classes and the test sample. We have termed the samples used for comparison to be prototypical samples.

Hence, a new class is discriminated as thus:

$$\text{new class}, \hat{k} = \begin{cases} True, & \text{if} D_e \leq \tau_d \ \forall k \\ False, & \text{otherwise} \end{cases} \tag{2}$$

where τ_d is the discrimination threshold. A new sample is only considered to be in a new class if the Euclidean distance is larger than the discrimination threshold τ_d for all the classes. In our final implementation of the learning approach, we developed a multitask model with the aim of simultaneously achieving class identification and new category recognition. This is done by modifying our final model as in 2.

The model is optimized by a combination of the triplet loss and the cross-entropy loss for embedding and classification respectively. The final loss function is then given as

$$Loss = Loss_{triplet} + \sum_{i}^{N} y_i \cdot \log(h(x_i)) \tag{3}$$

where $h(x_i)$ is the output of the classification branch of the model. The first part of the loss function represents the triplet loss optimization while the second part represents optimization for classification.

B. Experiments on Hand Control Gestures (HCG) Dataset

We set up an experiment, where a user uses hand gestures to send teleoperation commands to a robot system. The classes of gestures that were recorded are (i) Next, (ii) Previous, (iii) Up, (iv) Down, and (v) Action. The purpose of the gestures is to send the robot arm to preprogrammed positions in the workspace. We also introduced 4 auxiliary classes that infer when the user is not performing any gesture including when the hand is in the scene. The auxiliary gestures are (i) Empty Scene, (ii) Hand flat but static, (iii) Hand clenched but static, and (iv) Hand open but static. Figure 3 shows samples of the hand control gestures. A total of 29,019 samples were collected and 20% was used for validation during training. The model is able to achieve up to 94% validation accuracy.

Figure 4 shows a result of the t-SNE [19] visualization of the dataset samples in the embedding space after training. It can be seen that the trained model is able to produce embeddings that allows the samples to be grouped in to categories in the embedding space.

Due to scarcity of point cloud gesture datasets, we also experimented our multitask learning approach on the MSR Action 3D dataset. Since the dataset has 20 classes, we used 15 classes for training of the multitask model using a simple 10 layer feedforward network, and the remaining 5 classes for testing the one-shot learning approach. The embedding space visualization is shown in Figure 5 where model is able to recognize the new categories as shown in the clustering of the unknown categories in the embedding space. It is also able to achieve a classification accuracy of 89.32% on the known categories.

III. Model Evaluation with Validation Rates and False Accept Rates

Given a pair of gestures, a L_2 distance $D(x_i, x_j)$ is used to determine the recognition of same or different categories. All gesture pairs i, j of the same categories are denoted with P_{same}, whereas all pairs of different categories are denoted with P_{diff}. Similar to [18], we define the set of all *true accepts* as

$$TA(d) = \{(i,j) \in P_{same}, with D(x_i, x_j) \leq d\} \tag{4}$$

These are the gesture pairs i, j that were correctly recognized as same at threshold d. Similarly

$$FA(d) = \{(i,j) \in P_{diff}, with D(x_i, x_j) \leq d\} \tag{5}$$

is the set of all pairs that was incorrectly recognized as same category (*false accept*). The validation rage $VAL(d)$ and the false accept rate $FAR(d)$ for a given threshold d are then defined as

$$VAL(d) = \frac{TA(d)}{P_{same}}, FAR(d) = \frac{FA(d)}{P_{diff}} \tag{6}$$

From Figure 6 through 9, we show the result of the variation of validation rates and false accept rates with a chosen threshold. In all the cases, a threshold of 1.0 seem

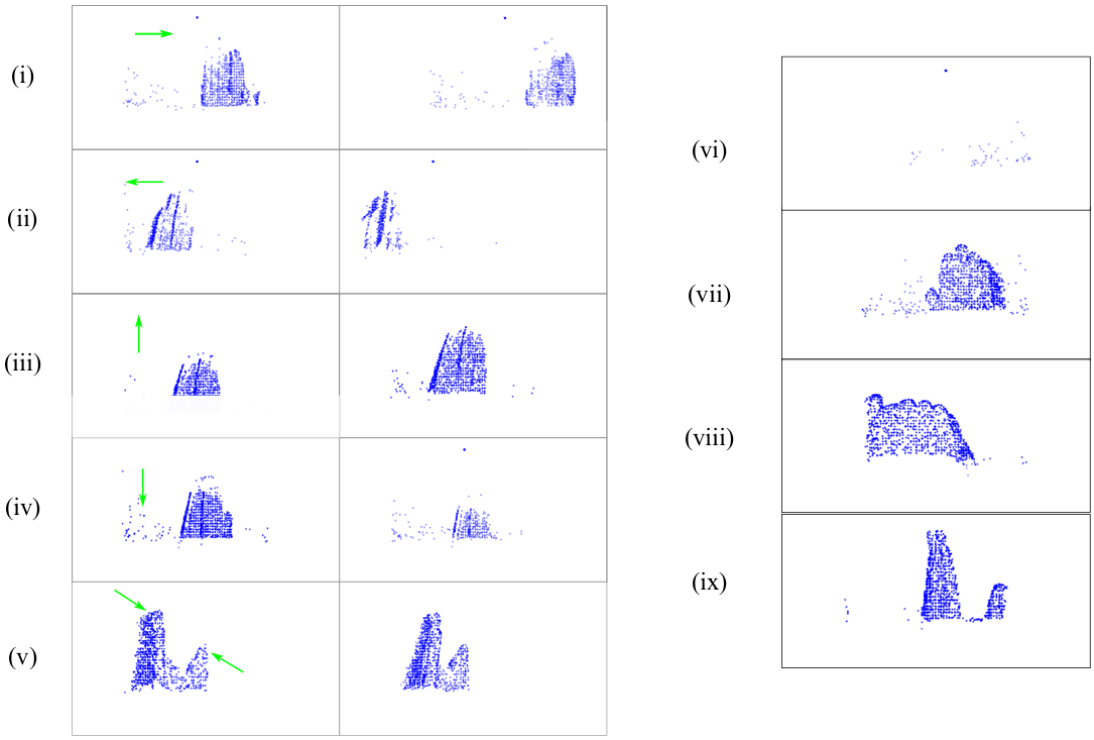

Fig. 3: Classes of hand gestures trained from point cloud data (i) Next, (ii) Previous, (iii) Up, (iv) Down, (v) Action/Grasp. The auxiliary classes are (vi) Empty Scene, (vii) Hand flat but static, (viii) Hand clenched but static, and (ix) Hand open but static.

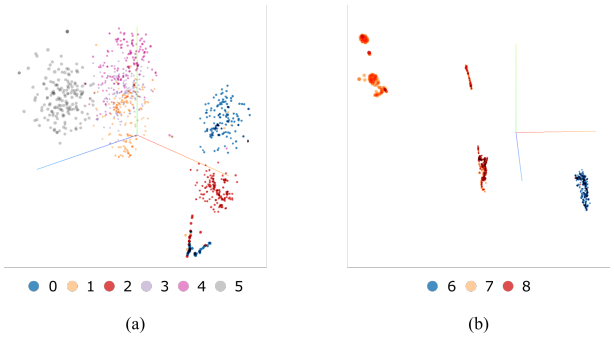

Fig. 4: Feature visualization of the class embeddings on the HCG dataset. (a) Clustering of 6 classes used for learning the embedding space. (b) Clustering of 3 left out classes to test the result of the embedding space.

to give the best trade of between high validation rate and lowest false accept rates. This is actually expected as the loss function enforces a distance between non-similar categories and it only makes sense to choose the mid point as the threshold for differentiating categories.

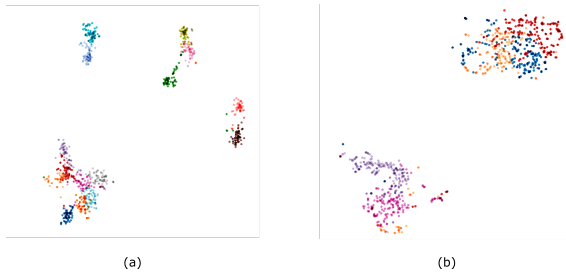

Fig. 5: Feature visualization of the class embeddings on the MSR Action 3D dataset. (a) Clustering of 15 classes used for learning the embedding space. (b) Clustering of 5 left out classes to test the result of the embedding space.

A. Comparison of Multitask Model with Normal One-shot approach

We compared the evaluation from the multitask model with a normal one shot approach where only the triplet loss is used. We found that using the multitask learning helped achieve high trade-off between *Validation Rates* and *False Accept Rates*. This is shown in Figures 8 and 9 where both models where tested on known and new classes of the MSR Action 3D dataset respectively. We see that the multitask model maintained a low *False Accept Rate* in most of the

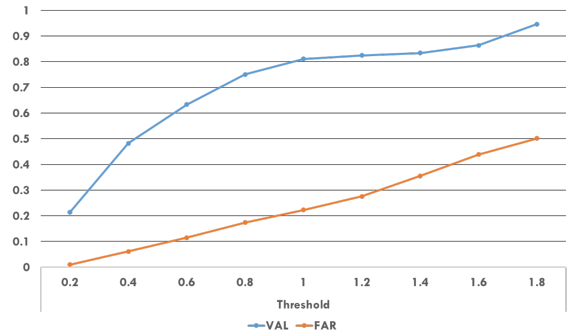

Fig. 6: Variation of Validation Rates and False Accept Rates with Threshold for Known Categories in the hand gesture Dataset.

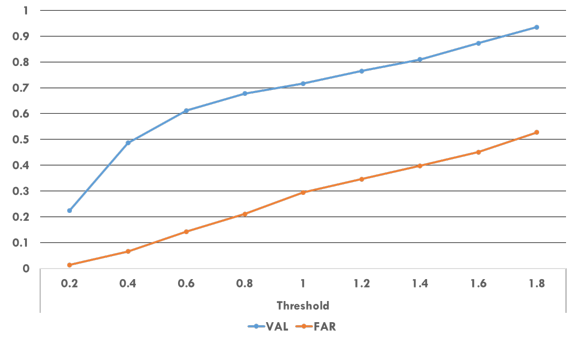

Fig. 7: Variation of Validation Rates and False Accept Rates with Threshold for unknown Categories in the hand gesture Dataset.

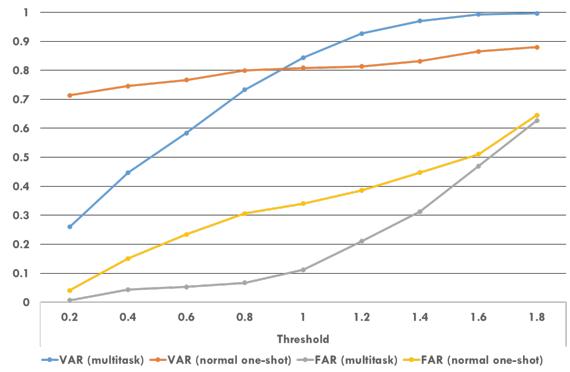

Fig. 8: Comparison of the Multitask Model and the Normal One-shot Model on Known Classes. The multitask model gives a better trade-off between validation rates and false accept rates by maintaining low false accept rates for most of the threshold values.

threshold values tested more than the normal one-shot model.

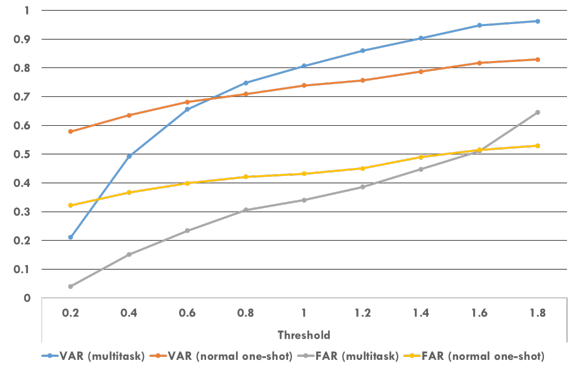

Fig. 9: Comparison of the Multitask Model and the Normal One-shot Model in New Classes. The trends are similar to those in Figure 8, however with higher False Accept Rates.

B. *Limitations of the Approach*

The following are the limitations that we have identified with this proposed approach.

1) The model begins to fail when the problem domain becomes too large. We note that this learning approach works well for face recognition [18] because the features on faces can be fairly standardized but dynamic gestures are more challenging because they have a higher degree of freedom, hence it will be very difficult to capture all the features to effectively categorize a gesture class using the 3D data.
2) Large amount of dataset is still needed to learn the generalization of classes in a problem domain. However, we propose that if a sufficiently large data is used for the initial embedding learning, the model can easily generalize to new classes.
3) Generalization into new categories is largely affected by the quality of original datasets. The dataset need to be highly representative of expected samples or future variations. This is a fundamental problem with most clustering approaches.

IV. Conclusion

In this paper we are able to show a new approach to rapidly learn new gesture classes from point clouds data. We used the approach called discriminative recognition to learn new classes by learning embeddings that can be discriminated in an Euclidean space. Using the results from class discriminations, we can differentiate a new class from one that has already been learned. This also helps us to recognize other samples of the newly discovered class. This paper contributes to the possibility of learning gesture classes without collecting a large number of dataset for new categories, or having to train a new model or retrain the old model to accommodate the new categories.

V. Acknowledgment

This work is partially supported by JSPS Grant-in-Aid 16H0653.

References

[1] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. 2016.

[2] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D Manning, and Andrew Y Ng. Zero-Shot Learning Through Cross-Modal Transfer. *NIPS*, pages 935–943, 2013.

[3] Matthias Bauer, Mateo Rojas-Carulla, Jakub Bartłomiej Świątkowski, Bernhard Schölkopf, and Richard E. Turner. Discriminative k-shot learning using probabilistic models. 1(Nips):3–6, 2017.

[4] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

[6] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese Neural Networks for One-shot Image Recognition. *International Conference on Machine Learning*, pages 1–8, 2015.

[7] Hanqin Tian, Chaoqun Lu, Jia Yang, Kamaljit Banger, Deborah N. Huntzinger, Christopher R. Schwalm, Anna M. Michalak, Robert Cook, Philippe Ciais, Daniel Hayes, Maoyi Huang, Akihiko Ito, Atul K. Jain, Huimin Lei, Jiafu Mao, Shufen Pan, Wilfred M. Post, Shushi Peng, Benjamin Poulter, Wei Ren, Daniel Ricciuto, Kevin Schaefer, Xiaoying Shi, Bo Tao, Weile Wang, Yaxing Wei, Qichun Yang, Bowen Zhang, and Ning Zeng. One-shot Learning with Memory-Augmented Neural Networks. *Global Biogeochemical Cycles*, 29(6):775–792, 2015.

[8] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical Networks for Few-shot Learning. (Nips), 2017.

[9] Franco Scarselli, Marco Gori, Markus Hagenbuchner, and Gabriele Monfardini. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[10] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing Machines. pages 1–26, 2014.

[11] Jia Lin, Xiaogang Ruan, Naigong Yu, and Ruoyan Wei. One-shot learning gesture recognition based on improved 3D SMoSIFT feature descriptor from RGB-D videos. *Proceedings of the 2015 27th Chinese Control and Decision Conference, CCDC 2015*, pages 4911–4916, 2015.

[12] Upal Mahbub, Hafiz Imtiaz, Tonmoy Roy, Md Shafiur Rahman, and Md Atiqur Rahman Ahad. A template matching approach of one-shot-learning gesture recognition. *Pattern Recognition Letters*, 34(15):1780–1788, 2013.

[13] Jun Wan, G Guo, and S Li. Explore Efficient Local Features from RGB-D Data for One-Shot Learning Gesture Recognition.

[14] Xiaojie Li and Shiyin Qin. One-shot Learning Gesture Recognition Based on Evolution of Discrimination with Successive Memory. *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*, pages 263–269, 2018.

[15] Joshua Owoyemi and Koichi Hashimoto. Learning Human Motion Intention with 3D Convolutional Neural Network. In *2017 IEEE International Conference on Mechatronics and Automation (ICMA*, pages 1810–1815, 2017.

[16] Cherdsak Kingkan and Joshua Owoyemi. Point Attention Network for Gesture Recognition Using Point Cloud Data. *2018, 29th British Machine Vision Conference*, pages 1–13, 2018.

[17] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[18] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:815–823, 2015.

[19] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research 9 (2008) 2579-2605*, 9:2579–2605, 2008.