A Prototype-Based Generalized Zero-Shot Learning Framework for Hand Gesture Recognition

Jinting Wu[1,2], Yujia Zhang[1] and Xiaoguang Zhao[1]

[1]Institute of Automation, Chinese Academy of Sciences, [2]University of Chinese Academy of Sciences

wujinting2016@ia.ac.cn

Abstract

- The task of Generalized Zero-Shot Learning (GZSL) for hand gesture recognition aims to recognize gestures from both seen and unseen classes by leveraging semantic representations.
- We propose an end-to-end prototype-based GZSL framework for hand gesture recognition which consists of two branches to tackle this challenge.
- We establish a hand gesture dataset that specifically targets this GZSL task, and comprehensive experiments on this dataset demonstrate the effectiveness of our proposed approach on recognizing both seen and unseen gestures.

Motivation

- Most existing works can only recognize a limited number of categories that have been seen during training.
- GZSL provides a solution for tackling the above challenges. However, GZSL approaches for dynamic hand gesture recognition are less explored.
- The recognition accuracy of existing zero-shot gesture recognition methods is not satisfactory enough.

Method

- **Overview of the Proposed Framework**
  - → The Prototype-Based Detector (PBD) learns a detector that determines whether an input sample belongs to a seen or unseen category, and meanwhile produces feature representations of unseen data.
  - → The zero-shot label predictor takes these features as input, and outputs predictions of samples from unseen classes through a learned mapping mechanism from feature to semantic space.
  - → These two branches are jointly trained in an end-to-end manner.

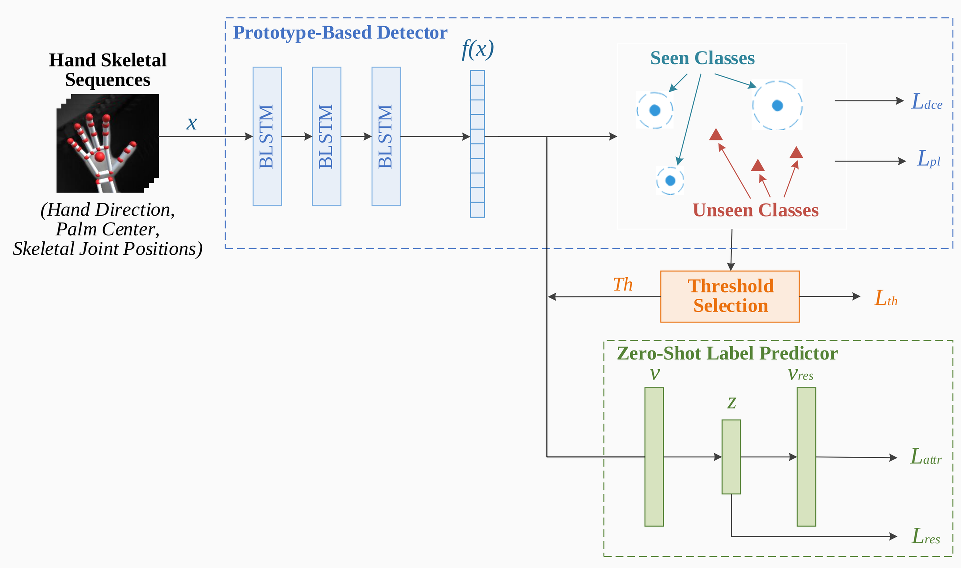

Fig. 1: Overview of the proposed framework

- **Prototype-Based Detector (PBD)**
  - → Using a multi-layer Bidirectional Long Short-Term Memory Networks (BLSTM) to extract temporal features
  - → Learning a fixed number of prototypes for each class
  - → The parameters of BLSTM and the prototypes are jointly trained through the distance-based cross entropy (DCE) loss L_{dce} and prototype loss L_{pl}:

$$L_{dce}((x,y)|\theta,M) = -\log\sum_{j=1}^{K}\frac{e^{-\gamma dis(p_{pbd}(x),m_{yj})}}{\sum_{k=1}^{C}\sum_{l=1}^{K}e^{-\gamma dis(p_{pbd}(x),m_{kl})}}, \quad (1)$$

$$L_{pl}((x,y)|\theta,M) = \|p_{pbd}(x) - m_{yj}\|_2^2. \quad (2)$$

- **Zero-Shot Label Predictor**
  - → Using a multi-layer Semantic Auto-Encoder (SAE) to predict the unseen gestures
  - → The loss function of SAE consists of an attribute loss L_{attr} and a reconstruction loss L_{res}:

$$L_{attr}((x,z_s)|\theta,\phi) = \|z - z_s\|_2^2, \quad (3)$$

$$L_{res}((x,z_s)|\theta,\phi) = \|v - v_{res}\|_2^2. \quad (4)$$

- **End-to-End Learning Objective**
  - → The above two branches can be jointly trained in an end-to-end manner.
  - → The joint learning objective of our end-to-end framework can be formulated as:

$$L((x,y,z_s)|\theta,M,\phi) = L_{dce} + \lambda_1 L_{pl} + \lambda_2 L_{attr} + \lambda_3 L_{res}. \quad (5)$$

- **Label Prediction**
  - → The model distinguishes the seen and unseen categories by comparing the minimum distance $d_m(x)$ in the prototype space with the thresholds $Th(x)$.
  - → Seen categories: the result given by the PBD module $\varepsilon(x)$
  - → Unseen categories: the result given by the SAE module $\varepsilon_u(x)$

$$d_m(x) = \min_{i=1}^{C}\left(\min_{j=1}^{K}\|p_{pbd}(x) - m_{ij}\|_2^2\right), \quad (6)$$

$$label(x) = \begin{cases}\varepsilon(x), d_m(x) \le Th(x)\\ \varepsilon_u(x), d_m(x) > Th(x).\end{cases} \quad (7)$$

Dataset

- The dataset contains 16 seen gestures and 9 unseen gestures which are captured by a Leap Motion Controller.
- The information such as hand direction, palm center and skeletal joint positions on a single right hand is recorded
- We design 11 attributes including hand movement and finger bending states.

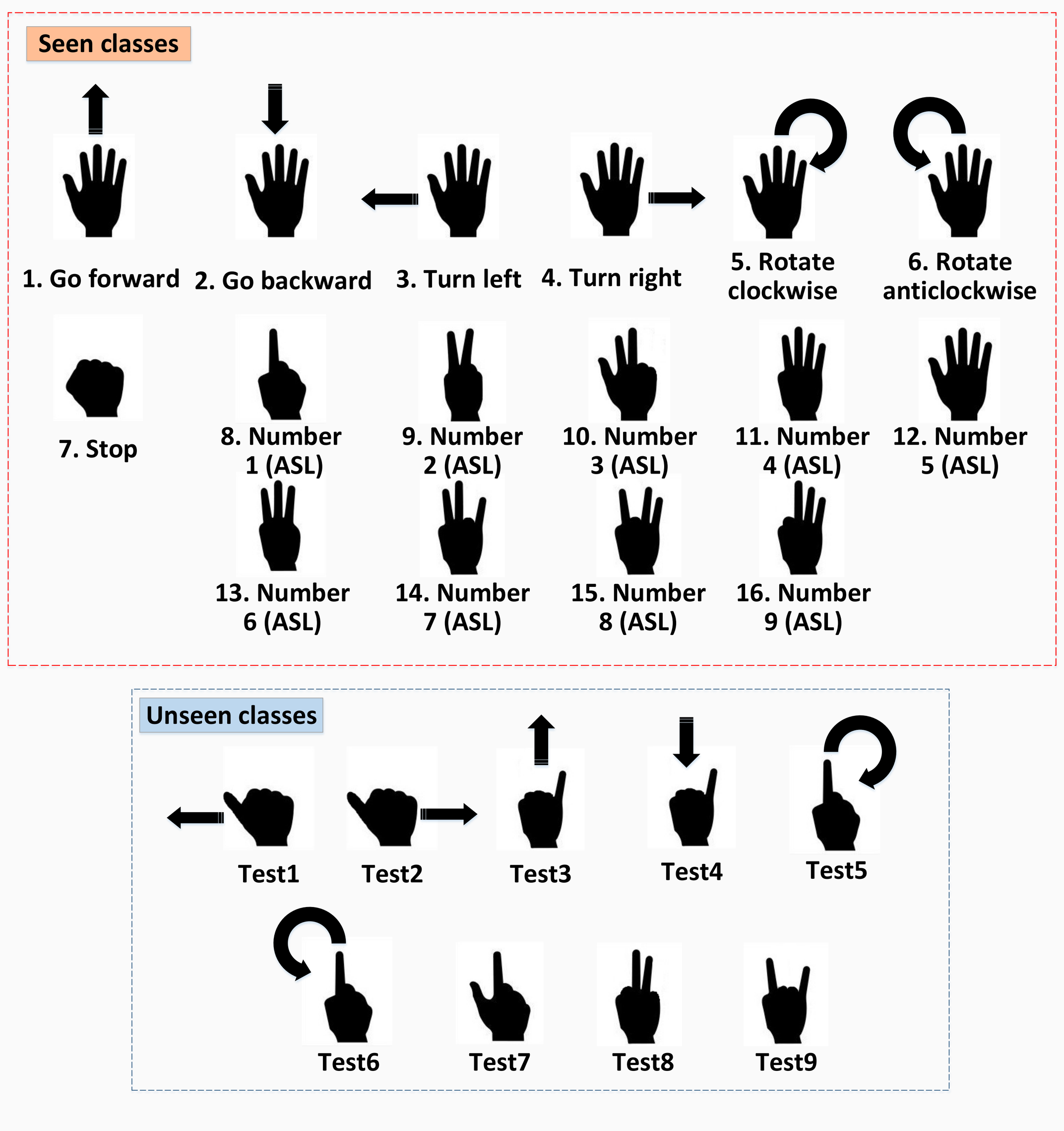

Fig. 2: Hand gesture categories

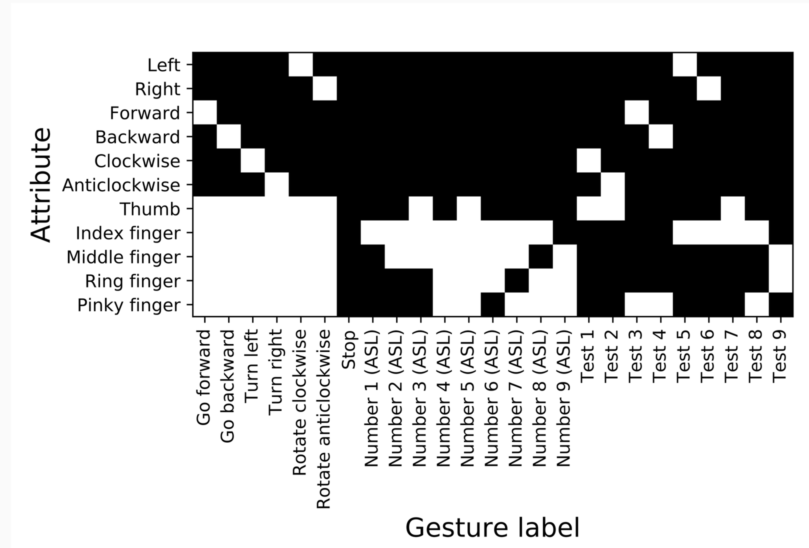

Fig. 3: Binary heat map of the categories and attributes

Experimental Results

- Evaluation Metrics
  - → The top-1 accuracy of seen classes and unseen classes: Acc_s and Acc_u
  - → Harmonic mean: H
- State-of-the-art Comparisons
  - → Zero-shot gesture recognition method: ESZSL [2]
  - → Generalized zero-shot object recognition method: CADA-VAE [3] and f-CLSWGAN [4]

Methods	Acc_s	Acc_u	H
ESZSL [2]	77.81%	13.89%	23.57%
CADA-VAE [3]	80.00%	53.89%	64.40%
f-CLSWGAN [4]	79.79%	55.00%	65.08%
Our Framework	**89.06%**	**58.33%**	**70.49%**

- Ablation Analysis
  - → The traditional SAE [1] without the prototype-based detector
  - → The framework with a fixed threshold for all seen categories
  - → The framework where two branches are trained separately

Methods	Acc_s	Acc_u	H	Test Time
SAE [1]	**91.88%**	15.00%	25.79%	0.023s
Fixed Threshold	84.69%	50.56%	63.31%	0.022s
PBD+SAE	90.63%	57.22%	70.15%	0.026s
Our Framework*	89.06%	**58.33%**	**70.49%**	**0.022s**

Conclusion

- We propose a prototype-based GZSL framework for hand gesture recognition. Two branches of our framework are introduced: a prototype-based detector and a zero-shot label predictor.
- The experimental results demonstrate that the proposed framework achieves a significant improvement over the state-of-the-art methods.
- In future work, we aim to extend this framework to a larger scale of gesture data in order to better support human-robot interaction in the real world.

References

[1] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3174–3183, 2017.

[2] Naveen Madapana and Juan Wachs. Zsgl: zero shot gestural learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 331–335, 2017.

[3] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8247–8255, 2019.

[4] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018.