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

Jinting Wu\(^1\), Yujia Zhang\(^1\) and Xiaoguang Zhao\(^1\)

\(^1\)Institute of Automation, Chinese Academy of Sciences, \(^2\)University of Chinese Academy of Sciences

wu.jinting2016@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 samples.
  - 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.

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 \( L_{p} \): 
  \[
  L_{dce}(x, y; \theta, M) = \sum_{i=1}^{N} \log \frac{e^{-d_\theta(x, y)}}{e^{-d_\theta(x, y)} + e^{-d_\theta(x, m_i)}} ,
  \]
  \[
  L_{p}(x, y; \theta, M) = \left\| \tilde{m}_y - m_y \right\|_2^2 ,
  \]

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 \( L_{att} \) and a reconstruction loss \( L_{rec} \):
  \[
  L_{att}(x, z|\theta, M) = \left\| x - z \right\|_2^2 ,
  \]
  \[
  L_{rec}(x, \hat{x}|\theta, M) = \left\| x - \hat{x} \right\|_2^2 ,
  \]

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:
  \[
  L((x, y); \theta, M) = L_{dce} + \lambda_1 L_{p} + \lambda_2 L_{att} + \lambda_3 L_{rec} .
  \]

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.

Legend

- Hand gesture recognition framework with two branches separating seen and unseen classes.

Experimental Results

- Evaluation Metrics
  - 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 \([1]\)
    - Generalized zero-shot object recognition method: CADASAE, CADASAE+SAE and F-CLSWGAN \([2] \)

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