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

Jinting Wu
Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Outline

- Introduction
- Methods
- Results
- Conclusion
Motivation

Most existing works can only recognize a limited number of categories that have been seen during training.

Generalized Zero-Shot Learning (GZSL) provides a solution for tackling the above challenges. However, GZSL approaches for dynamic hand gesture recognition are less explored.
Contributions

- We propose an end-to-end prototype-based GZSL framework for hand gesture recognition which consists of two branches.
- We establish a hand gesture dataset that specifically targets this GZSL task.
Outline

- Introduction
- Methods
- Results
- Conclusion
Methods

- Overview of the Proposed Framework
  - Two branches
  - Jointly training

![Diagram showing the proposed framework with two branches and joint training process.](image)
Methods

- **Prototype-Based Detector (PBD)**
  - Learning prototypes for each class
  - Distance-based cross entropy loss and prototype loss

\[
L_{dce} \left( (x, y) \mid \theta, M \right) = -\log \sum_{j=1}^{K} \frac{e^{-\gamma \text{dis}(p_{pbd}(x), m_{yj})}}{\sum_{k=1}^{C} \sum_{l=1}^{K} e^{-\gamma \text{dis}(p_{pbd}(x), m_{kl})}}
\]

\[
L_{pl} \left( (x, y) \mid \theta, M \right) = \left\| p_{pbd}(x) - m_{yj} \right\|_2^2
\]
Methods

- Zero-Shot Label Predictor
  - Using a multi-layer Semantic Auto-Encoder (SAE) to predict the unseen gestures
  - Attribute loss and reconstruction loss

\[
L_{\text{attr}} \left( (x, z_s) \mid \theta, \phi \right) = \left\| z - z_s \right\|_2^2
\]

\[
L_{\text{res}} \left( (x, z_s) \mid \theta, \phi \right) = \left\| v - v_{\text{res}} \right\|_2^2
\]
Methods:

- **End-to-End Learning Objective**
  \[ L((x, y, z_s)|\theta, M, \phi) = L_{dce} + \lambda_1 L_{pl} + \lambda_2 L_{attr} + \lambda_3 L_{attr} \]

- **Label Prediction**
  - Comparing the minimum distance in the prototype space \(d_m(x)\) with the thresholds \(Th(x)\).
  - **Seen categories:** PBD result \(\varepsilon(x)\)
  - **Unseen categories:** SAE result \(\varepsilon_u(x)\)

\[
label(x) = \begin{cases} 
\varepsilon(x), & d_m(x) \leq Th(x) \\
\varepsilon_u(x), & d_m(x) > Th(x) 
\end{cases}
\]
Outline

- Introduction
- Methods
- Results
- Conclusion
Results

- **Dataset**
  - 16 seen gestures and 9 unseen gestures
  - 11 attributes including hand movement and finger bending states
Results

- Experimental Results
  - State-of-the-art Comparisons
    - Zero-shot gesture recognition method: ESZSL\(^1\)
    - Generalized zero-shot object recognition method: CADA-VAE\(^2\) and f-CLSWGAN\(^3\)

<table>
<thead>
<tr>
<th>Methods</th>
<th>(Acc_s)</th>
<th>(Acc_u)</th>
<th>(H)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESZSL [15]</td>
<td>77.81%</td>
<td>13.89%</td>
<td>23.57%</td>
</tr>
<tr>
<td>CADA-VAE [11]</td>
<td>80.00%</td>
<td>53.89%</td>
<td>64.40%</td>
</tr>
<tr>
<td>f-CLSWGAN [12]</td>
<td>79.79%</td>
<td>55.00%</td>
<td>65.08%</td>
</tr>
<tr>
<td>End-to-End Framework (Ours)</td>
<td>\textbf{89.06%}</td>
<td>\textbf{58.33%}</td>
<td>\textbf{70.49%}</td>
</tr>
</tbody>
</table>

Results

Experimental Results

Ablation Analysis

- The traditional SAE\(^1\) without the prototype-based detector
- The framework with a fixed threshold
- The framework where two branches are trained separately

<table>
<thead>
<tr>
<th>Methods</th>
<th>(Acc_s)</th>
<th>(Acc_u)</th>
<th>(H)</th>
<th>Test Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLSTM+SAE [6]</td>
<td>91.88%</td>
<td>15.00%</td>
<td>25.79%</td>
<td>0.023s</td>
</tr>
<tr>
<td>End-to-End Framework (Fixed Threshold)</td>
<td>84.69%</td>
<td>50.56%</td>
<td>63.31%</td>
<td>0.022s</td>
</tr>
<tr>
<td>PBD+SAE</td>
<td>90.63%</td>
<td>57.22%</td>
<td>70.15%</td>
<td>0.026s</td>
</tr>
<tr>
<td>End-to-End Framework</td>
<td>89.06%</td>
<td>58.33%</td>
<td>70.49%</td>
<td>0.022s</td>
</tr>
</tbody>
</table>

Outline

- Introduction
- Methods
- Results
- Conclusion
Conclusion

- A prototype-based GZSL framework for hand gesture recognition
  - An end-to-end framework with two branches
  - A novel hand gesture dataset
  - Comprehensive experiments demonstrate the effectiveness of our proposed approach
Thanks for your attention!

- Jinting Wu
- Institute of Automation, Chinese Academy of Sciences; University of Chinese Academy of Sciences
- E-mail: wujinting2016@ia.ac.cn