Single-modal Incremental Terrain Clustering from Self-Supervised Audio-Visual Feature Learning

Reina Ishikawa, Ryo Hachiuma, Akiyoshi Kurobe, and Hideo Saito

{reina-ishikawa, ryo-hachiuma, kurobe.akiyoshi, hs}@keio.jp
Motivation

The key to an accurate understanding of terrain is to extract the informative features from the multi-modal data obtained from different devices

- RGB cameras
- depth sensors
- vibration sensors
- microphones
Problems

1. The data from multiple modal sensors are not always useful
2. The clustering model should update sequentially
3. Manual labeling is required
Single-modal Incremental Terrain Clustering from Self-Supervised Audio-Visual Feature Learning

1. A single-modal incremental terrain clustering framework learned in a self-supervised manner from audio-visual data
   - Combine an MVAE[1] for feature extraction and an IGMM[2] for cluster prediction
   - Clusters of terrains are updated during test-time

2. Input preprocessing
   - Generate edge image from visual data
   - Convert audio waveform into cochleogram

3. Evaluate the clustering accuracy and conduct extensive ablation studies

Multi-modal Variational Autoencoder

\[ ELBO(X) \equiv E_{q_{\phi}(Z|X)} \left[ \sum_{x_i \in X} \lambda_i \log p_\theta(x_i | z) \right] - \beta D_{KL}(q_{\phi}(z | x_i) \parallel p(z)) \]

- Product of Experts (PoE)
- sub-sampled training paradigm

\[ p_\theta : \text{encoder parameterized with } \theta \]
\[ q_{\phi} : \text{decoder parameterized with } \phi \]
\[ D_{KL}(q \parallel p) : \text{the Kullback–Leibler (KL) divergence between } p \text{ and } q \]
\[ \beta : \text{an annealing factor}^{[4]} \]

Methodology - Input preprocessing

- Raw image
  - Generate edge with Laplacian filter
- Audio raw waveform
  - Converted into cochleogram
Methodology - main scheme

\[ \mathcal{L} \equiv ELBO(x^{image}, x^{edge}, x^{audio}) + ELBO(x^{image}, x^{edge}) + ELBO(x^{audio}) + \beta D_{KL} \]
We tested our model on the dataset introduced by Kurobe et al. in [5]

- 21 movies
- 7 classes

Data splitting:

- 41315 training
- 7734 testing

Example visual data:
Qualitative Evaluation

Cluster 1 (Grass)
Cluster 2 (Pavement)
Cluster 3 (Carpet)
Cluster 4 (Linoleum)

Input test image
The training sample in the cluster
Input test audio
The training sample in the cluster
## Quantitative Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>NMI↑</th>
<th>ACC (%)↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>[5] w/o CNN</td>
<td>Audio+image</td>
<td>0.589</td>
<td>58.12</td>
</tr>
<tr>
<td>[5] w/ CNN</td>
<td>Image</td>
<td>0.001</td>
<td>23.18</td>
</tr>
<tr>
<td>Ours w/o update</td>
<td>Image</td>
<td>0.401</td>
<td>48.90</td>
</tr>
<tr>
<td>Ours w/ update</td>
<td>Image</td>
<td>0.377</td>
<td>50.63</td>
</tr>
<tr>
<td>Ours w/o update</td>
<td>Audio</td>
<td>0.353</td>
<td>50.30</td>
</tr>
<tr>
<td>Ours w/ update</td>
<td>Audio</td>
<td>0.500</td>
<td>74.39</td>
</tr>
</tbody>
</table>

## Ablation Study on Sound Input

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>w/o update</th>
<th></th>
<th>w/ update</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>NMI</td>
<td>ACC(%)</td>
<td>NMI</td>
<td>ACC(%)</td>
</tr>
<tr>
<td>MFCCs</td>
<td>Audio</td>
<td>0.559</td>
<td>55.92</td>
<td>0.235</td>
<td>34.73</td>
</tr>
<tr>
<td>MFCCs + cochleogram</td>
<td>Audio</td>
<td>0.389</td>
<td>49.57</td>
<td>0.443</td>
<td>47.98</td>
</tr>
<tr>
<td>Ours (cochleogram)</td>
<td>Audio</td>
<td>0.401</td>
<td>48.90</td>
<td>0.377</td>
<td>50.63</td>
</tr>
<tr>
<td>MFCCs</td>
<td>Image</td>
<td>0.295</td>
<td>47.26</td>
<td>0.389</td>
<td>61.02</td>
</tr>
<tr>
<td>MFCCs + cochleogram</td>
<td>Image</td>
<td>0.318</td>
<td>45.72</td>
<td>0.423</td>
<td>67.71</td>
</tr>
<tr>
<td>Ours (cochleogram)</td>
<td>Image</td>
<td><strong>0.353</strong></td>
<td><strong>50.30</strong></td>
<td>0.500</td>
<td><strong>74.39</strong></td>
</tr>
</tbody>
</table>
Single-modal Incremental Terrain Clustering from Self-Supervised Audio-Visual Feature Learning

Reina Ishikawa, Ryo Hachiuma, Akiyoshi Kurobe, and Hideo Saito

{reina-ishikawa, ryo-hachiuma, kurobe.akiyoshi, hs}@keio.jp