Abstract

We present a novel self-supervised framework for terrain type clustering using audio-visual data. Our method enables the terrain type clustering even if one of the modalities (either image or audio) is missing at the test-time.

Motivation

The key to an accurate understanding of terrain is to extract the informative features from the multi-modal data obtained from different devices, such as RGB cameras, depth sensors.

Problems:
1. The data from multiple modal sensors are not always useful, e.g., color is not useful for low illumination environment, and audio is not useful for noisy environment
2. The clustering model should update sequentially
3. Manual labeling is required

Contributions

- 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[3] for cluster prediction
  - Clusters of terrains are updated during test-time.
- Input preprocessing
  - Generate edge image from visual data
  - Convert audio waveform into cochleogram
- Evaluate the clustering accuracy and conduct extensive ablation studies

Dataset

We used a dataset introduced in [4], which includes 21 independent movies with seven classes.

Dataset Splitting:
- Training set : 41315
- Testing set : 7734

Method

Input preprocessing

- Audio Data: Convert to 2D cochleogram
- Image Data: generate edge image using laplacian filater

Training: Multimodal-feature learning

The training is conducted multi-modal manner to extract the informative feature using the paradigm of MVAE [1]. We encode three modalities (i.e., RGB, edge image, and audio).

The loss function:

\[
L = ELBO(x^{\text{image}}, x^{\text{edge}}, x^{\text{audio}}) + ELBO(x^{\text{image}}, x^{\text{edge}}) + ELBO(x^{\text{audio}}) + \beta_{\text{rel}}
\]

\[
\beta = \text{an annealing factor}[2]
\]

Testing: Single-Modal Incremental Clustering

Our method can predict the terrain cluster from a single modality data (either image or audio).

Even though the unseen terrain type appears, IGMM assigns a new cluster index.

Results

Qualitative Evaluation

Our method successfully predict the correct terrain cluster index from a single modality data.

Quantitative Evaluation

The incremental update at the test time improves the clustering accuracy.

Ablation Study on Visual Input

Taking RGB+Edge input outperformed only taking RGB input.

Ablation Study on Sound Input

Cochleogram method is the most suitable for preprocessing among the three preprocessing methods.

References