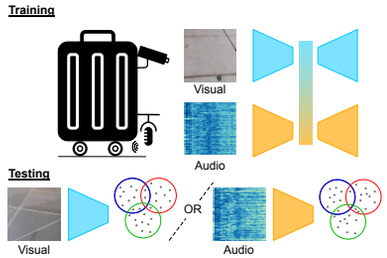
Keio University

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

Abstract

We present a novel self-supervised framework for terrain type clustering using audio-visul 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

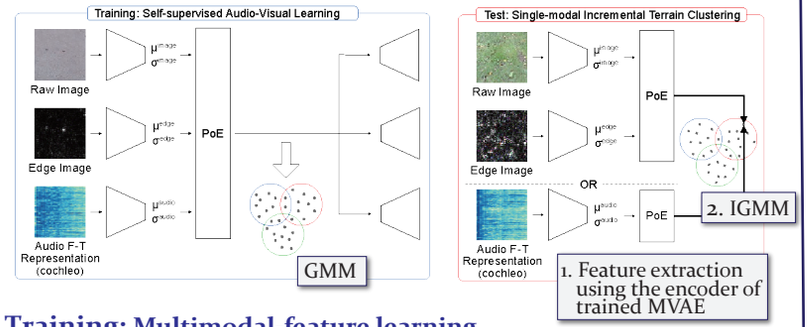
Contributions

- 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

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:

$$\mathcal{L} \equiv ELBO(x^{image}, x^{edge}, x^{audio}) + ELBO(x^{image}, x^{edge}) + ELBO(x^{audio}) + \beta D_{KL}$$

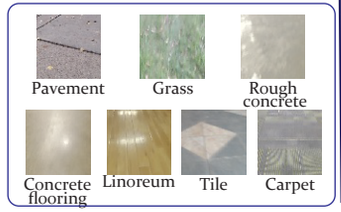
β : an annealing factor [2]

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.

Dataset

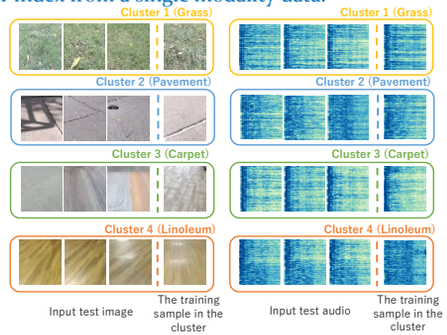
We used a dataset introduced in [4], which includes 21 independent movies with seven classes.
Dataset Splitting:
- Training set : 41315
- Testing set : 7734



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.

Method	Input	NMI ↑	ACC (%) ↑
[4] w/o CNN	Audio+image	0.589	58.12
[4] w/ CNN	Image	0.001	23.18
Ours w/o update	Image	**0.401**	48.90
Ours w/ update	Image	0.377	**50.63**
Ours w/o update	Audio	0.353	50.30
Ours w/ update	Audio	**0.500**	**74.39**

Ablation Study on Visual Input

Taking RGB+Edge input outperformed only taking RGB input.

Input	w/o update		w/ update	
	NMI ↑	ACC(%) ↑	NMI ↑	ACC(%) ↑
RGB	0.272	40.16	0.389	57.53
Ours (RGB+Edge)	**0.353**	**50.30**	**0.500**	**74.39**

Ablation Study on Sound Input

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

Method	Input	w/o update		w/ update	
		NMI	ACC(%)	NMI	ACC(%)
MFCCs	Audio	**0.559**	**55.92**	0.235	34.73
MFCCs + cochleogram	Audio	0.320	0.389	**0.443**	49.98
Ours (cochleogram)	Audio	0.401	48.90	0.377	**50.63**
MFCCs	Image	0.295	47.26	0.389	61.02
MFCCs + cochleogram	Image	0.318	45.72	0.423	67.71
Ours (cochleogram)	Image	**0.353**	**50.30**	**0.500**	**74.39**

References

[1] M. Wu and N. Goodman, "Multimodal generative models for scalableweakly-supervised learning," inInternational Conference on NeuralInformation Processing Systems, 2018, pp. 5580—-5590
[2] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick,S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual conceptswith a constrained variational framework," inInternational Conferenceon Learning Representations, 2016.
[3] P. Engel and M. Heinen, "Incremental learning of multivariate gaussianmixture models," inAdvances in Artificial Intelligence, 2010, pp. 82–91.
[4] A. Kurobe, Y. Nakajima, H. Saito, and K. Kitani, "Audio-visual self-supervised terrain type discovery for mobile platforms," CoRR, vol.abs/2010.06318, 2020