Keio University

CALAMVS GLADIO FORTIOR

1858

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

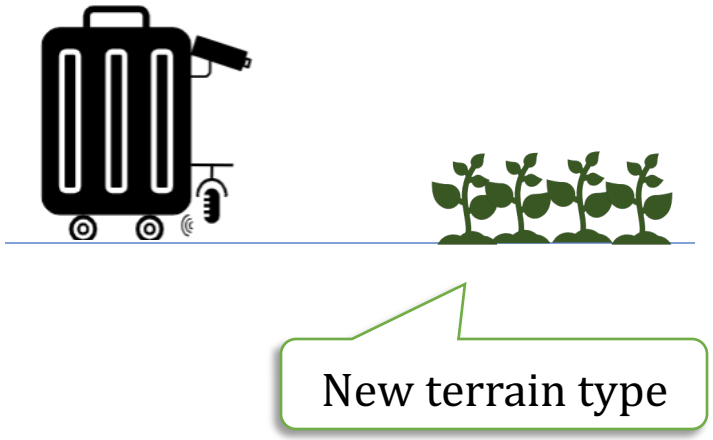
The key to an accurate understanding of terrain is to extract the informative features from the multi-modal data obtained from different devices

- RGB cameras
- depth sensors
- vibration sensors
- microphones

Problems

1. The data from multiple modal sensors are not always useful

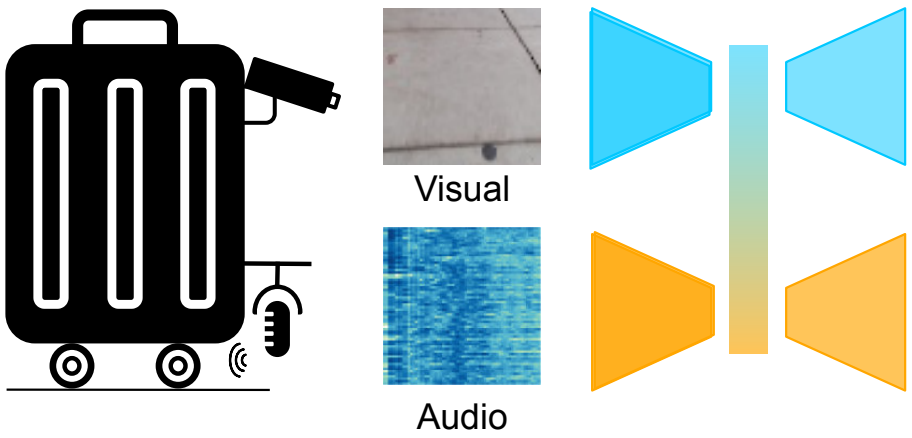
2. The clustering model should update sequentially



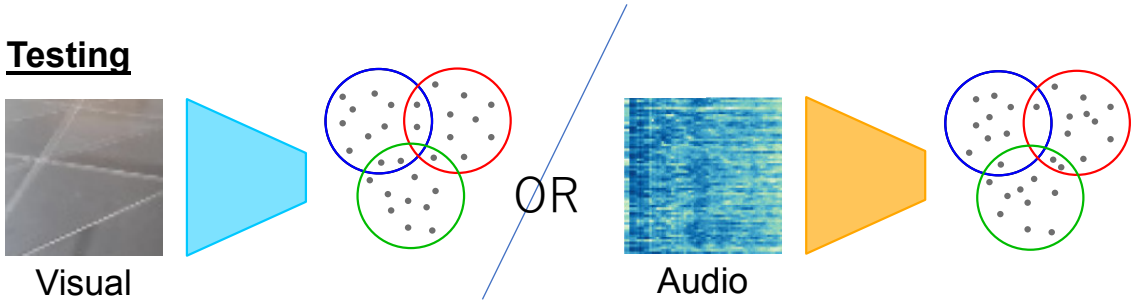
3. Manual labeling is required

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

Training



Testing



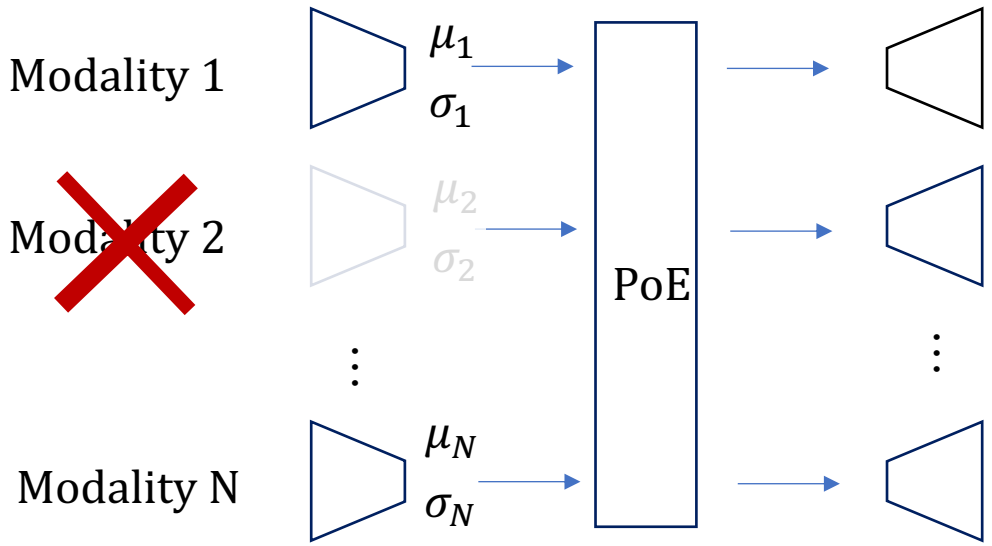
1. A single-modal incremental terrain clustering framework learned in a self-supervised manner from audio-visual data
   - Combine an MVAE[1] for feature extraction and an IGMM[2] for cluster prediction
   - Clusters of terrains are updated during test-time

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

3. Evaluate the clustering accuracy and conduct extensive ablation studies

[1] M. Wu and N. Goodman, "Multimodal generative models for scalableweakly-supervised learning," inInternational Conference on NeuralInformation Processing Systems, 2018, pp. 5580—-5590
[2] P. Engel and M. Heinen, "Incremental learning of multivariate gaussianmixture models," inAdvances in Artificial Intelligence, 2010, pp. 82–91.

Multi-modal Variational Autoencoder



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

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

p_θ : *encoder parameterized with* θ

q_ϕ : decoder parameteraized with ϕ

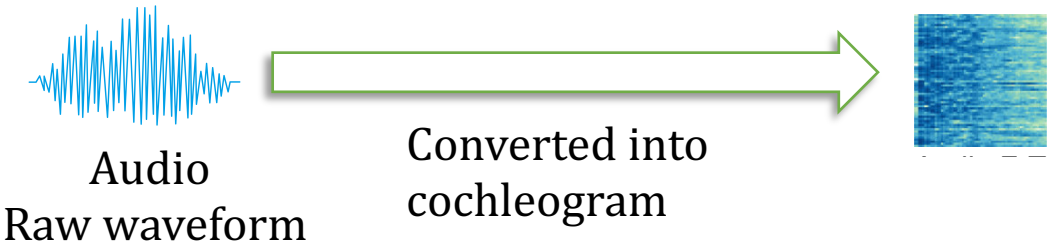
$D_{KL}(q \parallel p)$: *the Kullback − Leibler (KL) divergence between p and q*

β : *an annealing factor*[4]

[3] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in International Conference on Neural Information Processing Systems, 2018, pp. 5580—-5590.

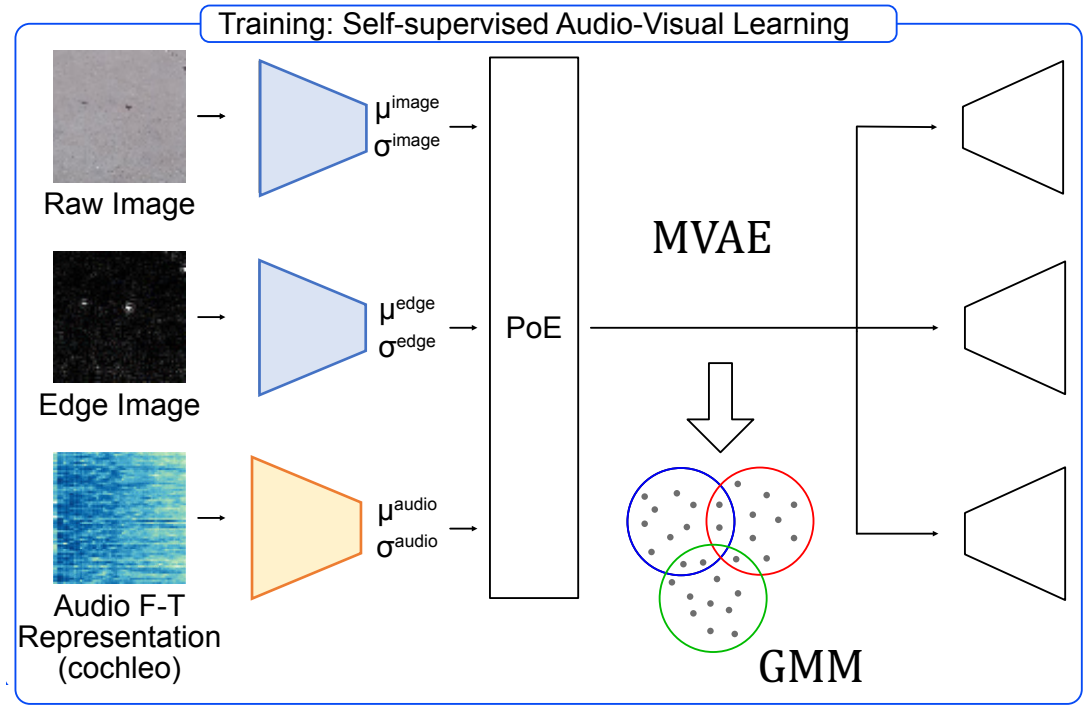
[4] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick,S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual conceptswith a constrained variational framework," in International Conference on Learning Representations, 2016.

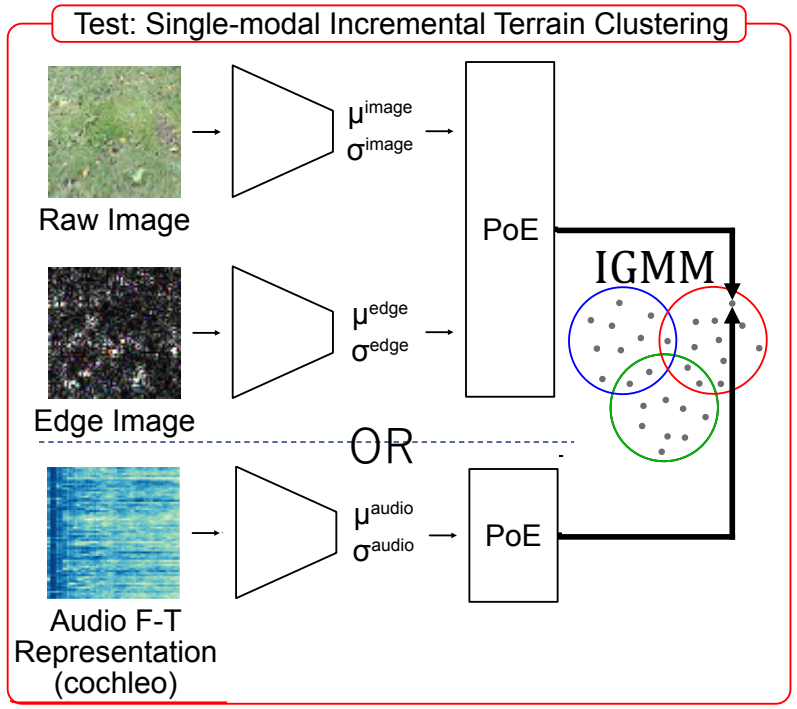
Methodology -Input preprocessing



Methodology - main scheme

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





Dataset

We tested our model on the dataset introduced by Kurobe et al. in [5]

- 21 movies
- 7 classes

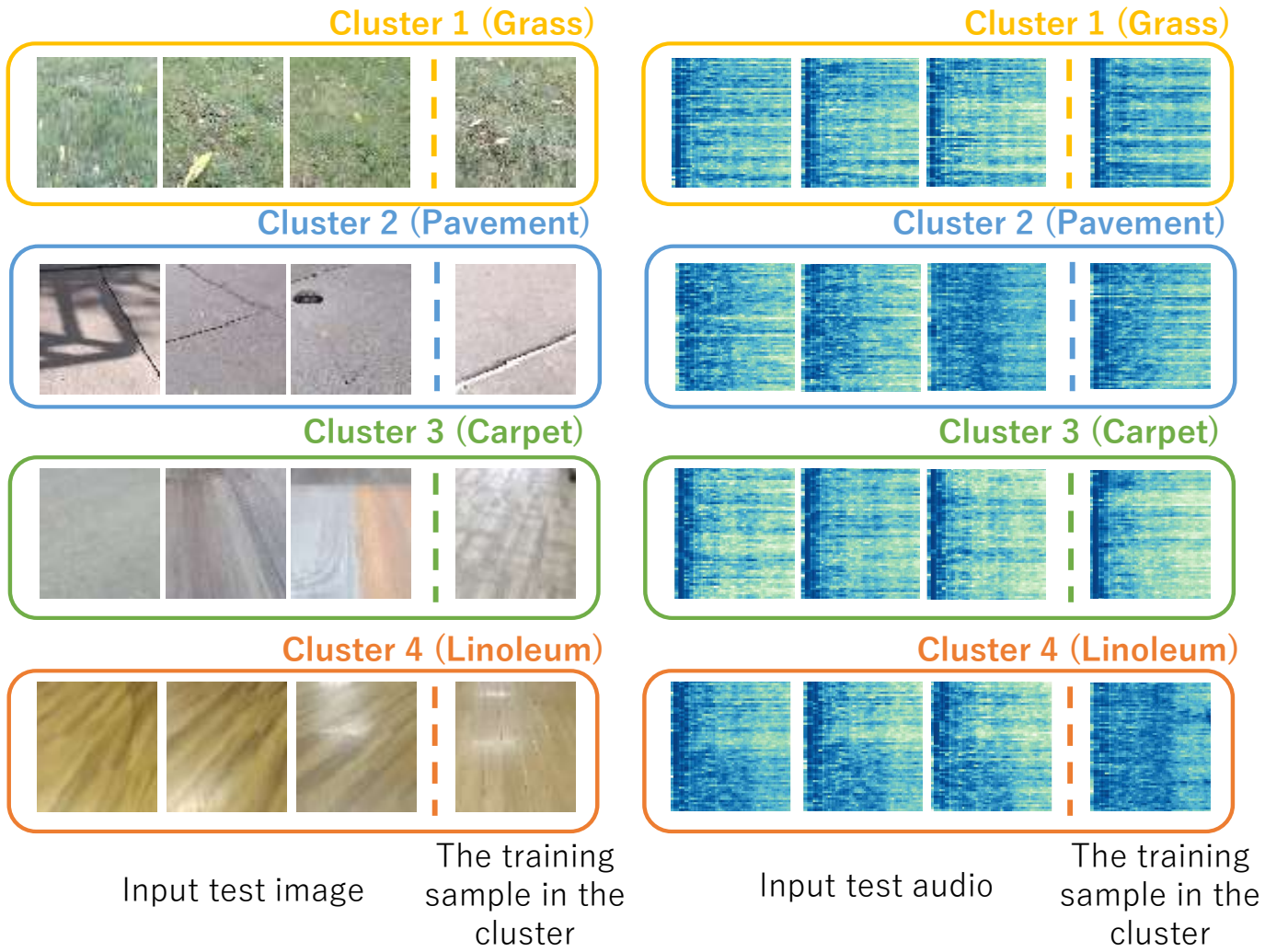
Data splitting:

- 41315 → training
- 7734 → testinig

Example visual data:

[5] 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

Qualitative Evaluation



Quantitative Evaluation

Method	Input	NMI ↑	ACC (%) ↑
[5] w/o CNN	Audio+image	0.589	58.12
[5] 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**

[5] 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

Ablation Study on Sound Input

Method	Input	w/o update		w/ update	
		NMI	ACC(%)	NMI	ACC(%)
MFCCs	Audio	**0.559**	**55.92**	0.235	34.73
MFCCs + cochleogram	Audio	0.389	49.57	**0.443**	47.98
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**

Keio University

CALAMVS GLADIO FORTIOR

1858

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