

自己教師あり学習を用いたシングルモーダルデータによる 地面材質のクラスタリング

石川玲奈^{1,a)} 八馬遼^{1,b)} 黒部聡亮^{1,c)} 斎藤英雄^{1,d)}

概要：地形を正確に把握するためには、さまざまなセンサーから得られるマルチモーダルデータから有益な特徴量を抽出することが重要である。RGB カメラやデプスセンサ、振動センサー、マイクロフォンなどがそのセンサーの例である。特にロボティクス分野では、シングルモーダル、あるいはマルチモーダルな手法を導入して特徴量の抽出に成功した論文は複数存在する。しかし実場面において、マイクが機能しないような混雑時、あるいは RGB カメラが機能しないような暗闇の中というような **extreme conditions** と呼ばれる場面にロボットは直面することがあり、既存研究においてはこのような場面が考慮されていないことが現状である。そこで本論文では、マルチモーダル変分オートエンコーダと混合ガウスモデルを用いて、画像データと音データを対象とした、地面の材質のクラスタリングのための新しいフレームワークを提案する。この手法はテスト時に画像または音声のいずれかのモダリティが欠落していてもクラスタリングを可能にする手法である。さらに我々は提案手法の有効性を示すため、従来のマルチモーダルな材質のクラスタリング手法を用いてクラスタリング精度を評価し、いくつかの **ablation study** を行った。

キーワード：自己教師あり学習、マルチモーダル学習、クラスタリング

1. はじめに

屋外においてカメラを用いて地面の材質や状態を把握することは、ロボット開発や自動車両制御のアプリケーションが広まっている現在、コンピュータ・ビジョンにおいて非常に注目を集めている研究分野である[2][3][4][5][6]。自律走行の分野においては[4]、ロボットの走行に悪影響を与える可能性がある地面の種類が存在することから、地面の種類の分類は非常に重要である。同様に、ロボット周辺の地面情報を取得することは、ロボットの自律的な経路修正に役立つ[3][6]。また、支援ロボット分野においては、視覚障害者に地面の種類に応じた危険を前もって警告することができる[7]。

地面の材質の視覚的認識は、たとえ同じ種類であっても、見た目が異なる場合があり、逆に、異なる種類であっても、見た目は非常に似通っている場合があるため、非常に困難を極める。そこで、視覚に基づく地面の分類における曖昧さに対処すべく、他のモダリティに基づく分類法、例えば音ベース[8][9][10]、触覚ベース[11][12]、あるいは、振動ベース[13]の手法が提案されている。音ベースの分類法では、ロボットの車輪と地面の接触音を拾うことで、地面の特徴を得る。これらの従来の研究では、それぞれのモダリティが地面の種類を識別するのに有効であることが証明されているが、単一のモダリティのみを用いる手法では、ノイズが含まれたり、建物やシーンの変化に対応できなかったりするため、曖昧さが残ってしまう。そこで、近年はマルチモーダル学習を行うことが、多くの学習タスクにおいて、互いのモダリティを補完し合い、より堅牢で優れた結果を導くことが実証されている[14]。本稿における提案手法は、

マルチモーダル学習の枠組みに従い、聴覚と視覚という 2 つのモダリティを活用して、堅牢に地面の材質を識別するための特徴を学習するものである。

マルチモーダルな地面のタイプ分類手法はいくつか提案されている[1][5][6]。Otsu ら[5]は画像と振動データから分類を行う手法を提案した。Zürn ら[6]は、ロボットのナビゲーションシステムのため、音声データから自己教師あり学習を行い、認識した地面に意味的なラベリングを行う手法を提案した。Kurobe ら[1]は、音と画像のマルチモーダルな自己教師あり学習を行い、抽出された特徴量を基に地面の種類のクラスタリングを行った。このクラスターラベルを用いて、画像ベースの畳み込みニューラルネットワーク (CNN) を学習し、地面の種類を推測するという手法である。しかしながら、実世界においてこのような手法を地面識別のアプリケーションとして展開するには、いくつかの課題点が浮上する。

第一の課題は、複数のモーダルセンサからのデータは必ずしも有用とは限らない、ということである。極端な条件のもとでは特定のセンサからは有益な特徴をとらえることが困難になり得るのである。例えば、RGB カメラの場合、夜間など照度が低い状態では、地面の質感を画像から読み取ることが困難になる[15]。また、マイクロフォンの場合、混雑している場所において、車両と地面の接触音をとらえることが難しくなる。しかし、従来のマルチモーダルな手法[1][5][6]では、画像のみからの予測など、単一の入力からの推論しかできなかった[1][6]。そこで本稿では、収集されるモダリティの一部が非有益である可能性を考慮する。具

¹ 慶應義塾大学
Department of Information and Computer Science, Keio University,
Yokohama 223-8522, Japan
a) reina-ishikawa@keio.jp

b) ryo-hachiuma@keio.jp
c) kurobe.akiyoshi@keio.jp
d) hs@keio.jp

目的には、音または画像のいずれか一つのモダリティのみでテスト可能にする。なお、提案するフレームワークは、データが極端な環境条件のもとで収集されたかどうかを検出することそのものは目的としておらず、テスト時には一方のモダリティが予め排除されていることを前提としている。

第二の課題は、このモデルをロボットに適用した場合、ロボットはデータを逐次的に取得するため、ロボットがデータに遭遇すると、クラスタリングモデルはデータのクラスタリングインデックスを推測するだけでなく、推測モデルを徐々に更新していく必要がある、という点である。さらに、クラスタリングモデルが実行中に更新されると、学習時には見られなかった地面の種類をクラスタリングできる可能性があるため、有用性が高まると言える。

加えて、最先端の分類法の多くは、教師あり学習であり、膨大な量のデータ数を必要とするが、データサンプルにラベルを手動で割り当てる[16]必要があるため、非常に労力がかかるといえる課題もある。一方、自己教師あり学習では、異なる入力間の相互関係を利用して学習データのラベル付けを自動的に行うことができるため、手動でのラベルづけ作業を大幅に削減することができる[17]。

本稿では、地面材質のクラスタリングのための、マルチモーダル自己教師あり学習のフレームワークを提案する。このフレームワークは、テスト時に単一のモダリティを受け入れ、クラスタリングモデルは逐次的に更新される。筆者の知る限り、このようなモデルを実世界のアプリケーションに展開するため、上記の重要な課題を解決する研究は未だ成されていない。テスト時に欠落しているモダリティを処理するために、提案するフレームワークでは、マルチモーダル変分オートエンコーダ(MVAE) [18]を採用している。この手法は、欠損データに対してロバストな手法である。また、潜在変数は多変量ガウス分布からサンプリングされるため、本手法ではインクリメンタルガウス混合分布(IGMM)を用いて逐次的にクラスタリングを行う。さらに、地面の種類をクラスタリングに有益な特徴量をオーディオデータ、およびビジュアルデータから抽出するための入力データの前処理方法を紹介する。オーディオデータはコクログラムにより時間一周波数表現に変換され、ビジュアルデータはエッジ画像と、生の画像の両方をエンコードする。本論文では、提案手法を Kurobe らの手法 [1]を従来のマルチモーダルな地面のタイプのクラスタリングモデルとして、提案手法と比較評価し、さらに提案手法の有効性を示す幾つかの ablation study を行う。

以下に、本稿の主な貢献を示す：

- 本研究は、筆者が知る限り、マルチモーダルなオーディオ・ビジュアルデータから自己教師的に学習されたモデルから、単一モーダルによって逐次的に地面材質のクラスタリングフレームワークを提案する

最初の研究である。

- 提案手法は、MVAE と IGMM を組み合わせた地面材質種類のクラスタリング手法である。IGMM クラスタリングアルゴリズムを用いることで、逐次的にクラスタリングを行うことができ、テスト時にガウス混合モデルを更新することが可能になるのである。
- 材質のクラスタリングのため、情報量の多い潜在変数を生成するための入力データの前処理方法を示す。
- 従来のマルチモーダルな地面の種類をクラスタリング精度と比較評価し、提案手法の有効性を示すため、ablation study を実施する。

2. 関連研究

本研究は、MVAE モデルと IGMM を用いて、地面のクラスタリングを、マルチモーダルに自己教師あり学習によって行う点で、ユニークな位置づけにある。MVAE により、テスト時にはビジュアルデータまたはオーディオデータのどちらか一方でシングルモーダルな推論を行うことが可能である。

2.1 地面の材質識別手法

自律走行型ロボットのナビゲーションシステムの経路設計において[19][20]、地形の状態がロボットの安全・安定な走行に影響するため、その把握は必須であると言える。このサブセクションにおいて、従来の単一モーダル法とマルチモーダル法による材質のタイプ識別の方法を紹介する。

単一モーダルベース

地面の状態を理解するための手法は、多くの論文で提案されているが、その中には、オーディオ [8][9][10]、触覚 [11][12]、振動 [13]、そしてビジュアル [3] [21][22]を使用するものがある。

分類手法の中には、視覚ベースのものがある：ステレオベース[3]と、特徴ベース[22]、スペクトラルベース[21]のアプローチがある。しかしながら、視覚ベースの分類法は輝度や反射の影響を非常に受けやすい。最近の研究では、自己教師あり学習やマルチモーダル手法において、視覚データを他のセンサデータと一緒に使用することが一般的になっている。音を用いたクラスタリングは光の条件に影響を受けないため、ロボティクス分野、特に脚型ロボット[8]や車輪駆動型ロボット[9][10]での研究が盛んに行われている。

マルチモーダルベース

地面の種類を認識におけるマルチモーダル手法は、ここ数年で活発に研究されている[1][5][6][23]。Zürn らはオーディオビジュアルベースの自己教師あり学習による地面のタイプのセマンティックセグメンテーション手法を提案した[6]。モデルは自己教師あり学習であるが、画像特徴抽出器と音声特徴抽出器が完全に独立しているため、テスト時に

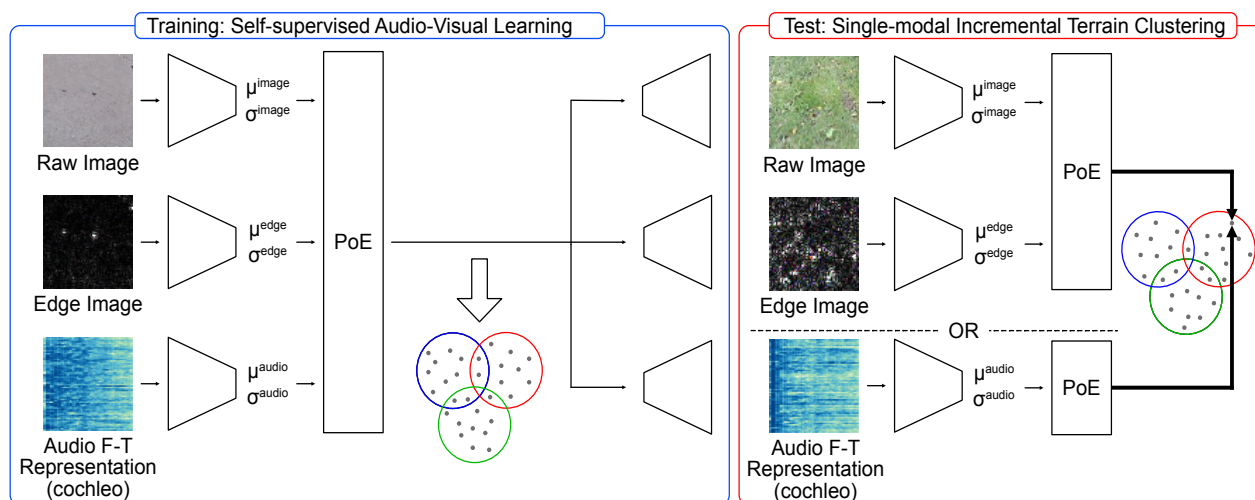


図 1 特徴抽出と、GMM を用いた特徴量クラスタリングの二つのモジュールで構成される。テスト時は、学習した MVAE のエンコーダ部分と、IGMM を用いたクラスタリングから構成される。提案手法では、マルチモーダルな学習を行う一方で、単一モダリティ(ビジュアルデータ又はオーディオデータ)からクラスタを推測する。

オーディオとビジュアルのどちらかが欠けている状況には対応できない。これに対し、提案手法では、そのような極端な状況が発生することを想定し、一つの強力に結合された構造を使用する。

黒部ら[1]は二つの独立した VAE を用いて、画像や音からの特徴量抽出した。彼らが提案する手法と提案手法とでは、目的の設定が異なる。Kurobe らの手法は、画像と音声の両方で学習したモデルを用いて、画像のみから地形の種類を識別することを最終的な目標としているため、学習の過程でカテゴリ数が固定されてしまっている。一方、本研究の最終目標は、識別ではなく、MVAE による地面の特徴のクラスタリングである。また、学習時に画像と音声の両方を用いることは、テスト時のクラスタリングの精度向上につながる。提案手法では、テスト環境に応じて、全てのセンシングデータをシングルモーダルにテスト時に利用したり、他のセンシングデータと組み合わせてマルチモーダルにテスト時に利用したりすることができることが特徴である。

Natalia ら[23]はジェスチャーや近距離行動認識のためのマルチモーダル融合戦略を提案した。彼らの提案した random dropping of separate channels (ModDrop) は、モデルが任意のモダリティの組み合わせを得ることができるという点で、提案手法と似ている研究である。

提案手法は、上記の手法と同様に、マルチモーダルな情報を活用するために、ビジュアルデータとオーディオデータの両方から学習する。従来研究においては、学習時にオーディオ・ビジュアルデータから情報量の多い特徴を学習することを目的としており、ビジュアルデータのみから材質の種類を推論している。本研究でも、オーディオデータから情報量の多い特徴量を抽出することを目的としているの

だが、テスト時には画像と音声のどちらかのデータからクラスタリングを行うことで、実環境での極端な撮影条件にも対応できるようにしている。

2.2 自己教師あり学習

マルチモーダルデータによって、あるモダリティを使って別のモダリティの学習をスーパーバイズするアプローチが多くあります。Wellhausen ら[24]は RGB ベースのセマンティックセグメンテーションを完全自動化することによるナビゲーションシステムを提案した。

Books ら[25]の手法はサポートベクタマシン (SVM) を用いた振動ベースの分類法を採用し、ビジュアルベースの分類をスーパーバイズする。

Zürn ら と Kurobe ら[6][1]はビジュアルデータと音オーディオデータを関連づけることで自己教師あり学習を行い、限られたカテゴリ数の中で、地面の材質のクラスタリングを行うことに成功している。本稿で提案する手法は、マルチモーダルな学習により、単一モダリティによるテスト結果が、別のモダリティによって改善されるという点で、自己教師型ととらえることができる。

2.3 オーディオデータの前処理方法

44100Hz 又は 48000Hz は、デジタルオーディオデータの主要なサンプルレートであるが、ノイズに強くするためオーディオ入力次元数は非常に大きくなる可能性がある。この問題を解決するために、データを前処理する研究がある。例えば、Libby ら[9]は、生の音信号の前処理として、時間領域と周波数領域の分析を行った。Ojeda ら[10]は離散的フーリエ変換 (STFT) ベースの log スケールスペクトログラムを音の前処理に用いた。小さなウィンドウサイズで

は地形の特徴を捉えることは困難であるため、提案手法では2.8秒の音のセグメントを64の時間領域のピースに分割し、その各ピースを64の周波数領域のピースに分割することで、時間-周波数領域とする手法を採用した。

3. 提案手法

本章では、図1のフレームワークの構成を説明する。セクションエラー! 参照元が見つかりません。では、入力された音声及び画像データの前処理を行い、セクション3.3ではMVAEを使用した自己教師あり学習による特徴量学習の学習方法を、そしてセクション3.4ではIGMMを使用した特徴量のクラスタリング手法を紹介する。さらに、セクション3.2では、MVAEの前段階であるVAEについての説明を行う。テスト時には、単一のモダリティ(画像又は音)にアクセスし、VAEのエンコーダを用いて特徴量ベクトルを抽出しGMMモデルを逐次的に更新していく。

3.1 入力データの前処理方法 ビジュアルデータ

VAEを用いて画像を潜在ベクトルにエンコードするために、本フレームワークでは、RGB画像に加えてエッジ画像をとり、潜在ベクトルにエッジ情報を明示的に含める。エッジ情報は地面のタイプをクラスタリングするための重要な手掛かりであるため、この問題は地面の種類のクラスタリングというタスクにおいて非常に重要である。例えばエッジの情報がないと、灰色のTileの画像と灰色のCarpetの画像を区別することは困難である。提案手法では、エッジ画像をRGB画像とは別に明示的にエンコードする。具体的には、ノイズ除去されたエッジ画像を生成するためにラプラシアンフィルタを採用する。本手法では、生のRGB画像に対して、 5×5 のカーネルを持つガウシアンフィルタとのカーネルを持つ4近傍ラプラシアンフィルタを適用する。

オーディオデータ

1次元の生のオーディオデータを二次元のデータに変換してからニューラルネットワーク(NN)に入力されることがある[28]。現在多くのオーディオ変換の手法が提案されており、中でもメル周波数ケプストラム(MFCCs)は音の特徴量の最も有効な表現方法の一つとして用いられてきた。MFCCsは人間の聴覚特徴を模倣したケプストラムベースの特徴表現方法であり、ピッチ情報が省略される。しかし、地面のクラスタリング課題において、車輪が転がる音を考える時、ピッチ情報はロボットが走る床の材質に依存するため、重要な手がかりとなる。そこで、本稿では、もう一つの処理方法であるコクレオグラムを使用する。

コクレオグラムは、ガンマトーンを用いたフィルタリング手法である。人間の聴覚システムの原理をより詳細にモ

デル化することで、システムの性能を向上させることができるという仮説のもと、詳細な生物物理学的コクレオモデルからコクレオグラム画像を得る[29]。

3.2 変分オートエンコーダ

VAEは、エンコーダ p を θ でパラメータ化したデコーダ q を ϕ でパラメータ化したエンコーダの間のボトルネック層で、低次元の特徴を球形ガウス形式で抽出する強力な手法である。このアルゴリズムの肝は、学習時に最大化したいデータの周辺尤度の扱いにくさを解消するために、下境界(evidence lower bound, ELBO)を最適化する、ということである。ELBOの定義は以下ようになる:

$$ELBO(x) \equiv \mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] - \beta D_{KL}(q_{\phi}(z|x) \parallel p(z)). \quad (1)$$

ここで、 $D_{KL}(q \parallel p)$ は2つの分布 p と q の間のKL(Kullback-Leibler)ダイバージェンスであり、 β は D_{KL} がエポック初期に急激に増加して損失の減少が阻害される影響を緩和するための係数である[30]。 z は潜在ベクトルであり、 x を入力とすると、オートエンコーダの出力 $\mu(x)$ と $\sigma(x)$ が与えられ、VAEは $q(z|x)$ から z を直接サンプリングする代わりに、 $\epsilon \sim \mathcal{N}(0, I)$ をサンプリングすることで、 $\mathcal{N}(\mu(x), \sigma(x))$ から潜在ベクトル z をサンプリングする。

3.3 マルチモーダル特徴量学習

提案フレームワークは、MFCCとコクレオグラムを組み合わせて生成したRGBの生画像 x^{raw} 、エッジ画像 x^{edge} 、2次元の音声データの3つのモダリティを1つの潜在ベクトルにエンコードする。テスト時に欠落しているモダリティを処理するために、MVAE[18]を採用する。

WuとGoodmanはVAEを拡張してMVAE[18]を提案した。入力が一つの1つの x_1 ではなく、 N 個のモダリティ $x_1, x_2, \dots, x_N$ に拡張されていることが特徴である。このモデルは各モダリティが条件付き独立であると仮定して成立しているため、再構成されたモダリティは一つの共有された潜在ベクトル z からデコードすることができる。

MVAEは条件付き独立の仮定のもと、個々のエキスパートモデルの変分パラメータを組み合わせるエキスパート積(PoE)[31]を利用している。また、MVAEのサブサンプル・トレーニング・パラダイムにより、このモデルは数値的な問題を回避しつつ、モダリティの全てのサブセットが推論ネットワーク内で近い潜在変数を持つように強制することができる。

MVAEのELBOは以下のように定義される[18]:

$$ELBO(x) \equiv \mathbb{E}_{q_{\phi}(z|x)} \left[\sum_{x_i \in X} \lambda_i \log p_{\theta}(x_i|z) \right] - \beta D_{KL}(q_{\phi}(z|x) \parallel p(z)). \quad (2)$$

ここで、 λ_i は、各モダリティの損失への影響を調整するバ

ランスファクタである。さらに、MVAEにおいては、存在するモダリティのうち任意の集合 X から推論が可能であり、その目的関数は次のように書くことができる：

$$\mathcal{L} \equiv -ELBO(x_1, x_2, \dots, x_N) - \sum_{i=1}^N ELBO(x_i) - \sum_{j=1}^M ELBO(x_j). \quad (3)$$

ここで、 N はモダリティの数、 M は任意のサブセット X_j の数を表す。

提案手法では、画像とエッジをそれぞれ独立した MVAE の入力とし、それらを再構成するが、入力 x^{raw} と x^{edge} は根本的に同じであるため、本手法では画像とエッジを常に一つのユニットとみなしている。この一体化の副産物として、計算コストの削減がある。以上をまとめると、学習時における目的関数は次のように書くことができる：

$$\mathcal{L} \equiv -ELBO(x^{raw}, x^{edge}, x^{audio}) - ELBO(x^{raw}, x^{edge}) - ELBO(x^{audio}). \quad (4)$$

提案手法では、再構成誤差を表す式の第一項を計算するために、以下のように定義される平均 2 乗誤差 (MSE) を用いる：

$$MSE = \frac{1}{N} \sum_i^N \sum_k^D (y_i^{(k)} - x_i^{(k)})^2. \quad (5)$$

ここで、 y は再構成データ、 x は入力データ、 N は学習に用いるミニバッチのサイズ、 D は特徴量の次元である。

3.4 潜在ベクトルのクラスタリング

複数のデータが似た特性を持っている場合、潜在空間のガウス分布の中で確率的に互いに近く位置する傾向がある。一方、異なる特徴を持つデータは、互いに離れて配置される傾向がある。ここで重要なのは、地面の材質の分類が同じであっても、潜在変数が互いに離れてしまうことがあるということである。逆に、異なる材質（例えば、Carpet と Grass）でも色が似ていたり、ロボットが走行している時の音が似ていたりすると、時には潜在空間においてクラスターが重なってしまうこともある。本稿の目的にハードクラスタリングよりソフトクラスタリングが適しているのは、まさにこのためであり、GMM の確率的割当が適している理由でもある。

GMM を用いることのもう一つの本質的かつ強力な利点は、ユーザーがクラスタの数を厳密に決定する必要がないことである。この特性は、ラベルがない教師なし学習や、クラスタ数が増加しうる場合において非常に重要である。提案するフレームワークでは、Engel と Heinen[32]が提案した IGMM アルゴリズムを用いている。このアルゴリズムでは、最小尤度基準を用いて、各データをそれぞれ一つのクラスタに割り当てを行う。新しいデータポイント x は、尤度 $p(x|j)$ が最小尤度であるような既存クラスタ、あるいは新規クラスタ j に属することになる。パラメータ（すなわち、

平均 μ^j 、そして共分散 σ^j 、混合パラメータ $p(j)$ ）は事後確率 $p(j|x)$ の和として蓄積されていく。

4. 実験

4.1 データセット

提案手法の評価には、Kurobe らが[1]において導入したデータセットを使用する。このデータセットでは、超指向性マイクを用いてモノラルステレオの音データを、RGB カメラを用いて画像データを約 24fps で撮影している。

ただし、本研究においては、オリジナルのデータセットに変更を加えている。オリジナルのデータセットは、テスト時の RGB 画像のみから材質クラスを推論するもので、音データは含まれていない。また、Kurobe らは、シーンごとの評価しか行なっておらず、シーンを跨いだより一般的な性能評価はしていない。一方で本研究では、複数の異なるシーンに対して一般化することを目的としている。

データセットには、RGB フレームと音声データを含む 21 の独立した映像で構成されており、Pavement, Grass, Rough concrete, Concrete flooring, Carpet, Tile, Linoleum の 7 つのクラスが含まれている。また、RGB フレームとそれに対応する、オーディオクリップを生成するために、各画像フレームに対応する前後計 2.8 秒分を使用する。

超指向性マイクは車輪の音を拾い、RGB カメラは前方の地面を撮影しているため、オーディオクリップと対応する画像フレームの間には時間的なギャップがある。このギャップによる材質の境界の曖昧さを避けるため、提案手法ではカメラが境界を記録している 35 秒分を除外した。

最後に、41,315 の学習用データセットと 7,734 のテスト用データセットに分割した。なお、材質タイプのアノテーションを行ったが、これは評価のみを目的とし、学習時やテストのクラスタリング過程では使用されない。

4.2 ネットワーク構造

提案モデルの構造は、MVAE[18]で採用されている画像ネットワークに似ている。エンコーダは、カーネルサイズが 4×4 の 2 次元畳み込み層、バッチ正規化層、活性化関数 Swish[33]で構成されている。デコーダは、カーネルサイズが 4×4 の 2 次元の転置畳み込み層、バッチ正規化層、Swish で構成される。提案手法では、完全結合層をサイズ 5×5 の畳み込み層に置き換えることで、ネットワークのパラメータ数を削減している。

4.3 データの前処理

ビジュアルデータ

提案手法では、データセットの画像中央下部から 720×720 にクロッピングしたのち、 68×68 にサイズケーリングする。さらに、これらの画像を 64×64 の大きさにランダムにクロッピングすることで、データのオーグメンテーション

ンを行う。

オーディオデータ

音データは、セクション 3.1 で説明したコレオグラムで処理される。この時、漸近的フィルタ品質係数として 9.265、最小帯域幅として 24.7 を用いる。

4.4 学習の詳細

本稿の実験では、Adam optimizer[34]を採用する。学習率は 10^{-4} 、ミニバッチサイズは 300 に設定する。また、最初の 20 エポックで 0.0 から 1.0 までアニーリングするように設定する。

4.5 評価手法

本稿では、正規化相互情報量 (NMI) とクラスタリング精度 (ACC) の 2 つの評価指標を採用し、従来のクラスタリング手法[6]と比較する。

NMI については、以下のような式で表される：

$$NMI(Y, C) = \frac{2I(Y; C)}{H(Y) + H(C)}. \quad (6)$$

ここで、 $Y = \{y_1, y_2, \dots, y_K\}$ はクラスタラベルの集合、 $C = \{c_1, c_2, \dots, c_J\}$ はクラスタラベルの集合、 $H(x)$ は x 全体のエントロピーである。 $I(Y; C)$ は相互情報量で、次のように与えられる：

$$I(Y; C) = H(Y) - H(Y|C). \quad (7)$$

ここで、 $H(Y|C)$ は C の条件付きエントロピーである。この値が大きいくほど、NMI と精度の両方の指標で優れている。

また、クラスタリングの精度は以下のように定義される：

$$ACC(Y, C) = \sum_{k=1}^K \frac{\max_j |c_k \cap y_j|}{N}. \quad (8)$$

ここで、 N はデータポイントの総数、 y_k と c_j はそれぞれ Y と C の特定の成分を示している。

4.6 他の手法との比較

オーディオ・ビジュアルデータにより学習されたモデルによって、シングルモーダルに地面の材質のクラスタリングを行う手法は未だ提案されていない。そこで本稿は、本研究で用意したデータセットを用いて、[1]の手法と比較する。なお、Zürn らが提案した手法は、画像を材質ごとに分類することを目的としており、本研究の目的とは大きく異なるので、提案手法の結果とは比較を行わない。

本研究では、以下の二つの方法でクラスタリング結果を比較する：

- Kurobe らの手法[1]から CNN を除いたモデル

[1]は音や画像の入力に対して VAE を学習させ、学習したデータをクラスタリングして擬似ラベルを作成する。提案手法との公平な比較をするために、画像と音声の VAE を

学習し、抽出した特徴量を GMM でクラスタリングし、推論されたラベルをもとに NMI と ACC を算出する。

- Kurobe らの手法[1]から CNN を除かないモデル：

Kurobe らは音データと画像データの擬似ラベルを用いて自己教師的に学習した CNN を用いて、クラスタリングラベルを推論する。クラスタリングの結果得られた CNN の出力ラベルについては、真のラベルと出力ラベルを用いて NMI と ACC の値を算出している。

5. 実験結果

5.1 潜在特徴量の可視化

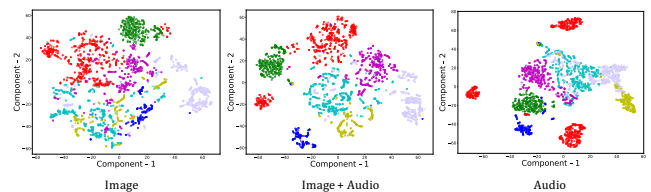


図 2 t-SNE による潜在ベクトルの可視化結果

図 2 はテストデータに対して t-SNE アルゴリズム[35]を用いて PoE により結合された潜在ベクトルを可視化したものである。この図では、材質の種類によって色分けされており、同じ色でラベル付けされたデータは、似たような特徴を持っている。この図から、クラスタはよく分類されており、真のクラスタラベルと高い相関関係にあることがわかる。

5.2 定性的評価

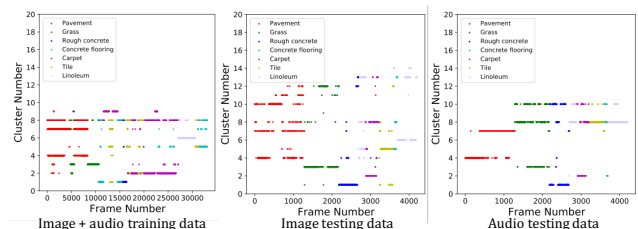


図 3 クラスタインデックスのプロット例

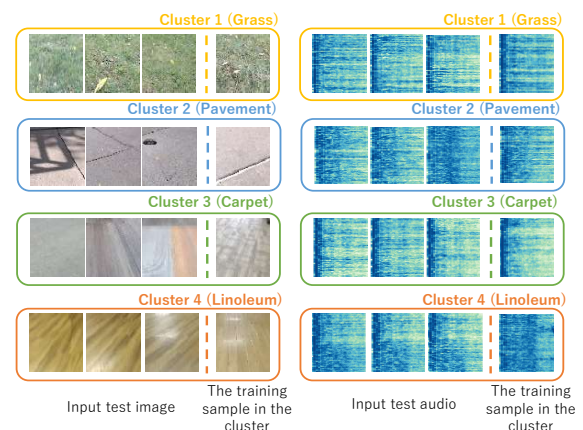


図 4 定性的評価の結果。同色の枠内のデータは同じクラスタに割り当てられたことを示している。

図 3 に提案手法の訂正的評価の結果を示す。各枠内の最右列は、同じクラスに属する学習データサンプルである。提案手法は、テスト時に画像または音の入力を受けて、クラスターのインデックスを推論する。この図から、提案手法はマルチモーダルな特徴量学習を行なっているにもかかわらず、単一のモダリティから正しいインデックスを推論することに成功していることがわかる。また、図 4 は割り当てられたクラスターインデックスをプロットした結果である。マルチモーダルに推論した学習データと、シングルモーダルに推論したテストデータのプロットを比較すると、ほとんどの場合、同じ真のラベルを持つデータが同じクラスに割り当てられていることがわかる。

5.3 定量的評価

表 1 Kurobe らの手法[1]との比較 (w/は with の略、w/o は without の略)。GMM の初期クラスターは 10 とした。

Method	Input	NMI ↑	ACC(%) ↑
[1] w/o CNN	Audio+image	0.589	58.12
[1] w/ CNN	Image	0.001	23.18
Proposed w/o update	Image	0.322	44.68
Proposed w/ update	Image	0.413	63.42
Proposed w/o update	Audio	0.278	36.50
Proposed w/ update	Audio	0.485	59.49

提案手法の定量的な評価を表 1 に示す。Kurobe らは材質の種類が多い大規模データを想定していないため、画像データのみを用いた場合には NMI や精度が急激に低下しているが、画像と音声の両方を入力した場合の NMI 値は提案手法を上回っている。また、表 1 では、逐次的アップデートを行った場合と行わなかった場合のクラスタリングの評価を示している。更新していないクラスタリングと更新したクラスタリングの精度を比較すると、テスト時に更新を行うことで、クラスタリングの精度が向上することがわかる。

5.4 ビジュアル入力の ablation study

表 2 RGB 画像とエッジ画像を使用した場合と、視覚情報の入力として RGB 画像のみを使用した場合の比較。学習にはビジュアルデータとオーディオデータの両方を利用し。GMM の混合成分の数は 10 に初期化される。

Method	Input	w/o update		w/ update	
		NMI ↑	ACC(%) ↑	NMI ↑	ACC(%) ↑
RGB	Audio	0.163	30.66	0.346	56.32
RGB+Edge	Audio	0.322	44.68	0.413	63.42

エッジ画像を入力に用いることの重要性についても評価した。表 2 よりエッジ画像を用いる手法と用いない手法を比

較すると、RGN とエッジの両方を入力する手法は、RGB のみを入力とする手法よりも、精度の面で 10% 上回ることがわかった。このことから、地形タイプのクラスタリングにエッジ情報を用いることが有効であることがわかる。

5.5 オーディオ入力の ablation study

表 3 RGB 画像とエッジ画像を使用した場合と、視覚情報の入力として RGB 画像のみを使用した場合の比較。学習にはビジュアルデータとオーディオデータの両方を利用し。GMM の混合成分の数は 10 に初期化される。

Method	Input	w/o update		w/ update	
		NMI ↑	ACC(%) ↑	NMI ↑	ACC(%) ↑
MFCCs	Audio	0.246	34.27	0.460	48.82
MFCCs+cochleogram	Audio	0.440	50.96	0.247	36.37
cochleogram	Audio	0.278	36.50	0.485	59.49

また、提案手法の有効性を確認するために、MFCCs のみを用いた特徴量表現、及びコクレオグラムと MFCCs を組み合わせた特徴量表現と比較する実験を行った。この結合特徴量は、MFCCs の特徴量 $x^{mfcc} \in \mathbb{R}^{16 \times 64}$ とコクレオグラムの特徴量 $x^{coch} \in \mathbb{R}^{48 \times 64}$ を連結することで、最終的な入力 $x^{con} \in \mathbb{R}^{16 \times 64}$ となる。

オーディオ処理の方法を比較した結果を表 3 に示す。MFCCs+cochleogram 法は、GMM を更新しない場合には cochleogram 法を上回っていたが、GMM を段階的に更新すると、NMI と ACC が cochleogram 法をおきく下回る。また MFCCs や MFCCs+cochleogram 法を音声処理に用いた場合、画像のみでテストした場合の精度は、音と画像の両方でテストした場合と同じくらい高いのだが、最初のクラスター数が増えるにつれて精度が致命的に低下する。これは、MFCCs や MFCCs+cochleogram で変換した音データが役に立たず、MVAE モデルは画像データに依存してしまうことを意味している。画像と音のバランスを考え、逐次的クラスタリングを考慮すると、3 つの音の前処理方法の中では、cochleogram が前処理に適していると言える。

6. 結論

本稿では、IGMM を用いて、MVAE の潜在変数から地面材質の種類をクラスタリングする新しいフレームワークを提案した。MVAE により、テスト時にビジュアルまたはオーディオデータのいずれかが利用できなくても、提案手法は機能し、その結果極端な状況下で片方のモダリティが欠損していても、モデルは部分的な入力を受け取ることができることが実証された。また、IGMM により逐次的に高性能なクラスタリングを実現した。加えてビジュアルデータとして RGB とエッジ画像を使用し、オーディオデータには cochleogram を用いた前処理を用いることで、材質のクラスタリングに有益な特徴量を抽出できることを実証した。

謝辞

この研究は JST (JPMJMI19B2)の支援によるものである。

参考文献

- 1) Akiyoshi Kurobe, Yoshikatsu Nakajima, Hideo Saito, and Kris Kitani. Audio-visual self-supervised terrain type discovery for mobile platforms. CoRR, abs/2010.06318, 2020.
- 2) A. Angelova, L. Matthies, D. Helmick, and P. Perona. Fast terrain classification using variable-length representation for autonomous navigation. In Conference on Computer Vision and Pattern Recognition, pages 1–8, 2007.
- 3) Roberto Manduchi, A. Castano, Ashit Talukder, and L. Matthies. Obstacle detection and terrain classification for autonomous off-road navigation. Autonomous Robots, 18:81–102, 01 2005.
- 4) D. F. Wolf, G. S. Sukhatme, D. Fox, and W. Burgard. Autonomous terrain mapping and classification using hidden markov models. In International Conference on Robotics and Automation, pages 2026–2031, 2005.
- 5) K. Otsu, M. Ono, T. J. Fuchs, I. Baldwin, and T. Kubota. Autonomous terrain classification with co- and self-training approach. Robotics and Automation Letters, 1(2):814–819, 2016.30
- 6) Jannik Zurn, Wolfram Burgard, and Abhinav Valada. Self-supervised visual terrain classification from unsupervised acoustic feature learning. CoRR, abs/1912.03227, 2019.
- 7) Muhammad Zeeshan Hashmi, Qaiser Riaz, Mehdi Hussain, and Muhammad Shahzad. What lies beneath one’s feet? terrain classification using inertial data of human walk. Applied Sciences, 9:3099, Jul. 2019.
- 8) J. Christie and N. Kottege. Acoustics based terrain classification for legged robots. In International Conference on Robotics and Automation, pages 3596–3603, 2016.89
- 9) J. Libby and A. J. Stentz. Using sound to classify vehicle-terrain interactions in outdoor environments. In International Conference on Robotics and Automation, pages 3559–3566, 2012.
- 10) Lauro Ojeda, Johann Borenstein, Gary Witus, and Robert Karlsen. Terrain characterization and classification with a mobile robot. Journal of Field Robotics, 23, Feb. 2006.
- 11) Kuniyuki Takahashi and Jethro Tan. Deep visuo-tactile learning: Estimation of material properties from images. CoRR, abs/1803.03435, 2018.
- 12) Byounggon Park, Jayoung Kim, and Jihong Lee. Terrain feature extraction and classification for mobile robots utilizing contact sensors on rough terrain. Procedia Engineering, 41:846 – 853, 2012.
- 13) C. Bai, J. Guo, and H. Zheng. Three-dimensional vibration-based terrain classification for mobile robots. IEEE Access, 7:63485–63492, 2019.
- 14) Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In International Conference on Computer Vision, Oct. 2019.
- 15) Di Hu, Lichao Mou, Qingzhong Wang, Junyu Gao, Yuansheng Hua, Dejing Dou, and Xiao Xiang Zhu. Ambient sound helps: Audiovisual crowd counting in extreme conditions. CoRR, abs/2005.07097, 2020.
- 16) K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.
- 17) Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. CoRR, abs/1812.05069, 2018.
- 18) Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In International Conference on Neural Information Processing Systems, pages 5580–5590, 2018.
- 19) M. Ono, T. J. Fuchs, A. Steffy, M. Maimone, and J. Yen. Risk-aware lane-tary rover operation: Autonomous terrain classification and path planning. In Aerospace Conference, pages 1–10, 2015.
- 20) A. Chilian and H. Hirschmüller. Stereo camera based navigation of mobile robot on rough terrain. In International Conference on Intelligent Robots and Systems, pages 4571–4576, 2009.
- 21) Xiuwen Liu and DeLiang Wang. Texture classification using spectral histograms. IEEE Transactions on Image Processing, 12(6):661–670, 2003.
- 22) P. Filitchkin and K. Byl. Feature-based terrain classification for little dog. In International Conference on Intelligent Robots and Systems, pages 1387–1392, 2012.
- 23) Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Mod-drop: adaptive multi-modal gesture recognition. CoRR, abs/1501.00102, 2015.
- 24) L. Wellhausen, A. Dosovitskiy, R. Ranftl, K. Walas, C. Cadena, and M. Hut-ter. Where should i walk? predicting terrain properties from images via self-supervised learning. IEEE Robotics and Automation Letters, 4(2):1509–1516, 2019.
- 25) C. Brooks and K. Iagnemma. Self-supervised classification for planetary rover terrain sensing. In Aerospace Conference, pages 1–9, 2007.
- 26) Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In Computer Vision and Pattern Recognition, pages 2536–2544, 2016.
- 27) Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In European Conference on Computer Vision, pages 649–666, 2016.
- 28) Di Hu, Lichao Mou, Qingzhong Wang, Junyu Gao, Yuansheng Hua, Dejing Dou, and Xiao Xiang Zhu. Ambient sound helps: Audiovisual crowd counting in extreme conditions. CoRR, abs/2005.07097, 2020.
- 29) Mladen Russo, Luka Kraljević, Maja Stella, and Marjan Sikora. Cochleogram-based approach for detecting perceived emotions in music. Information Processing & Management, 57(5):102270, 2020.
- 30) Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In International Conference on Learning Representations, 2016.
- 31) Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. Neural Computation, 14(8):1771–1800, 2002.
- 32) Paulo Engel and Milton Heinen. Incremental learning of multivariate gaussian mixture models. In Advances in Artificial Intelligence, pages 82–91, 2010.
- 33) Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. CoRR, abs/1710.05941, 2017.
- 34) Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015.
- 35) Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of machine learning research, 9:2579–2605, 2008.