マルチモーダル全方位画像の共起性と循環性を考慮した畳込み特徴学習による一般屋外環境識別

中嶋一斗　河村晃宏　○倉爪亮（九州大学）

1. はじめに

自律移動型ロボットにとって，周囲環境の状況を高精度かつ頑健に認識することは最も重要な課題の一つである．特に，走行環境を識別することで，障害物検出精度の向上や環境に適応的な行動計画など，コンテキストに沿った高信頼な視覚機能を実現することができる．例えば，人通りの多い住宅地を走行する自動車が，周囲環境を認識し，自律的に走行速度を下げるような環境適応型運転支援が可能となる．

環境種別を推定する一方法として，GPS で計測した自己位置の地理情報を利用することができるが，GPS の測定精度や地理情報と実環境との時間的・空間的ギャップが問題となる．そのため，ロボット本体に搭載した可視光カメラや赤外線カメラ，LiDAR を用いて，実際の動的環境を自律的に認識する技術開発が数多く行われている．一般的に可視光カメラが広く用いられるが，夜間や雨天時など日照の低下に伴い，シーンの見え方は劇的に変化してしまう．また，赤外線カメラ画像は，可視光カメラに比べて照明変化に頑健であるが，日射熱の影響を強く受ける．一方，LiDAR は周辺環境の幾何形状計測を主目的としているものの，レーザ照射に基づく原理上，屋外における日照条件の影響を受けにくい．近年では，全方位型 3D LiDAR が普及し，図 1 に示すような走行環境全方位の距離データおよび距離計測の副産物である反射強度データが得られる．反射強度はレーザ照射を受けた物体の材質に応じて変化するため，距離データの持つ幾何情報では表現できない詳細なテクスチャを含んでおり，“見え” 情報として環境認識タスクに活用できる可能性が高い．

こうした背景から，我々はこれまでに全方位型 3D LiDAR を利用した一般屋外環境識別手法の開発に着手し，6 種類の屋外環境を対象とした大規模 3 次元点群データセット Sparse Multi-modal Panoramic 3D Outdoor（Sparse-MPO）を構築した [1]．本稿では，全方位型 3D LiDAR の計測点群から生成される全方位距離画像と全方位反射強度画像を組み合わせた新たな一般屋外環境識別手法を提案する．特に，近年画像理解に関する諸タスクにおいて従来法よりも高い性能を示している畳込みニューラルネットワーク（Convolutional Neural Networks, CNN）を利用した End-to-end 学習モデルを構築し，Sparse-MPO を用いた性能評価により提案手法の有効性を示す．

2. 関連研究

環境を捉えた画像情報から一般環境種別を推定する手法は，これまでに数多く提案されているが，近年ではその多くが CNN をベースとしている．一般的な可視光画像を用いた取り組みとしては，Zhou ら [2] が屋内外の大規模データセットを構築し，代表的な CNN アーキテクチャを用いた識別性能評価を行っている．Song ら [3] は，可視光画像と距離画像が対となった大規模データセットを構築し，CNN による識別性能評価を行っている．しかし，各画素の可視光強度値と距離値が一対一に対応した画像を撮影できるセンサの制約上，識別対象は屋内環境に限られている．また，環境識別に関する多くの研究が単一方向の視覚情報に基づいており，環境特有の特徴を得られない可能性がある．

(a) 可視光画像

(b) 全方位距離画像（上）と全方位反射強度画像（下）

図 1: 市街地を撮影した可視光画像と同地点から LiDAR で計測した全方位画像の例 [1]

LiDAR から得られる 3 次元点群を対象とした CNN アーキテクチャとしては，Muturana ら [7] が，占有格子表現に変換した 3 次元点群を入力として，物体認識のための 3 次元畳込みニューラルネットワーク（3D CNN）を学習している．3D CNN は計測点群の幾何情報を厳密に保持できるが，2D CNN に比べて学習効率および計算効率が低いため，2 次元距離画像に変換し，CNN の多層化に取り組むアプローチが主流である．また，Shi ら [4] は，3 次元の物体 CAD モデルを円筒投影し，2 次元のパノラマ画像として CNN に入力した．さらに，対象物体の回転に不変な特徴を得るための Row-Wise Max Pooling（RWMP）層を提案している．循環構造を持った画像を扱う点で本稿の目的と関連が深い．

単一種の画像から特徴を抽出するのではなく，距離画像やマルチスペクトル画像のような異種データと組み合わせたクロスモーダル解析手法も数多く提案されている．代表的な手法は，多チャンネル画像として積層するもの [8] やモダリティ毎に複数の畳込み伝搬路を有する N-stream CNN [9] などである．近年では，モーダル毎に学習した CNN の推定確率を適応的に重み付けするアプローチが提案されている [5]．いずれも CNN 内部特徴表現の階層性に着目しているが，最適なアーキテクチャはパラメータ学習に利用されるデータの規模とタスクに依存している．

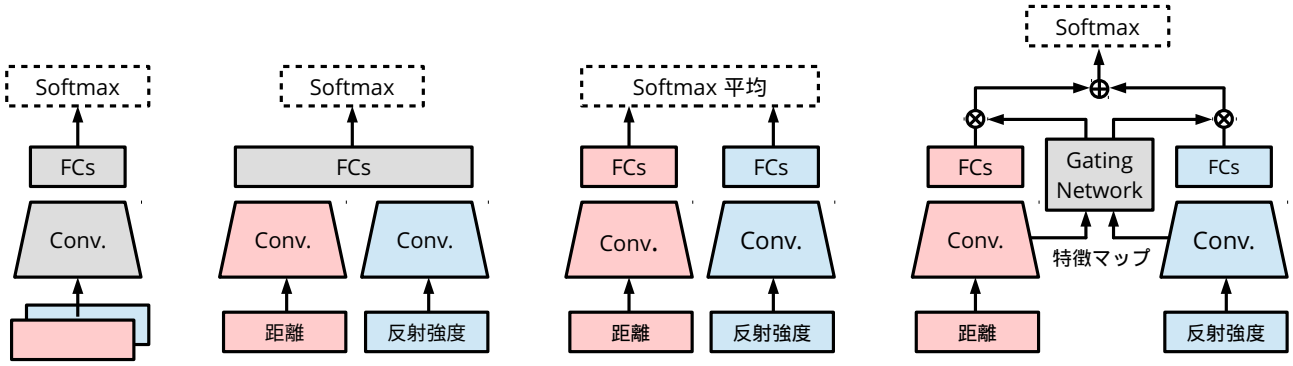


図 2: マルチモーダル全方位画像を入力とする CNN アーキテクチャ．Conv. は畳込み層，FCs は全結合層を表す．

3. 全方位画像による一般屋外環境識別

本稿では，全方位画像による一般屋外環境識別を対象として，以下の 2 点について議論する．

1. 全方位画像の循環性を考慮した畳込み特徴学習
2. 距離画像と反射強度画像に有効な特徴融合法

以降では，LiDAR 計測データから全方位画像を生成する手順，全方位画像特徴を学習するための CNN モデル，および全方位距離画像と全方位反射強度画像の融合アーキテクチャについて述べる．

3.1 入力データ

使用した全方位型 3D LiDAR（Velodyne HDL-32e）からは，スキャンライン上の計測点毎に距離値 $\in [0.0, 100.0]$ と反射強度値 $\in [0, 255]$ が得られる．通常，球面座標系にある計測点を直交座標系に変換することで 3 次元点群を得るが，円筒投影により計測点に対応する距離値と反射強度値を直接 2 次元平面に対応づけることで，図 1 の全方位画像を得る．Sparse-MPO に含まれる全方位画像は，垂直画素数が LiDAR のスキャンライン数 32，水平画素数が 1 ラインの計測点数 2166 となる．本研究では，計測データの欠損除去と CNN の学習効率の観点から水平画素数を 384 にダウンサンプリングした．

3.2 CNN による全方位画像特徴の学習

全方位距離画像と全方位反射強度画像のそれぞれを処理する CNN は，Simonyan ら [6] が ImageNet の識別実験に利用した VGG-11 モデルをベースとする．VGG-11 モデルは，8 つの畳込み層と 3 つの全結合層から構成され，全ての畳込みカーネルサイズが 3×3，プーリングサイズが 2×2 である．通常の畳込み層では，特徴マップの出力サイズ縮小を防ぐために，予め画像周囲を 0 で埋めるゼロパディングが多く用いられる．本稿では，全方位画像の循環構造を保持したまま特徴抽出を行うための循環畳込み層を提案し，VGG-11 モデルの 8 つの畳込み層全てを置き換える．循環畳込み層は，順伝搬時のパディング処理と逆伝搬時の勾配計算におけるデータの流出入を水平方向に循環させる．また，与えられる全方位画像の撮影方向（3D LiDAR の設置ヨー角）に不変な特徴表現を得るために，最終畳込み層のプーリング処理を Row-Wise Max Pooling (RWMP) [4] に置き換える．RWMP は，入力される特徴マップの行毎に最大値を出力することで，循環構造を持った画像の水平移動に不変性を与える．最終層を除く全ての畳込み層および全結合層の出力には，バッチ正規化 [11] を適用した．

3.3 距離情報と反射強度情報の融合

マルチモーダル全方位画像から屋外環境カテゴリを識別するための 4 つの CNN アーキテクチャについて述べる．図 2 に，各アーキテクチャの模式図を示した．

Early Fusion． 距離画像と反射強度画像をチャンネル方向に積層し，VGG モデルに入力する．両者の情報を最初の畳込み層で融合するため，入力画像空間上の局所的な共起性を学習する．

Late Fusion． VGG モデルの畳込み伝搬路を 2 つ設け，距離画像と反射強度画像を個別に与える．次にそれぞれの畳込み層出力を結合し，全結合層を通して確率分布を推定する．入力画像の空間情報が抽象化された意味的特徴を融合する．

Softmax Average． 距離画像と反射強度画像を個別に事前学習した 2 つの VGG モデルから確率分布を推定し，平均確率を算出する．

Adaptive Fusion． 距離画像 $\boldsymbol{x_d}$ と反射強度画像 $\boldsymbol{x_r}$ に対する中間特徴マップから，両者の信頼度 w_d, w_r を推定する Gating Network [5] を導入し，両モデルの推定確率を適応的に重み付けする．損失関数 L は次式で定義される．ただし，$f(\cdot)$ は CNN，$g(\cdot)$ は Gating Network，$r(\cdot)$ は CNN の中間出力を表す．また，N はミニバッチの大きさ，$\boldsymbol{y}$ は 1-of-K 表現した正解ラベルである．

$$L(\boldsymbol{y}, \boldsymbol{x}) = -\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{y}_k{}^T \log F(\boldsymbol{x}_k) \tag{1}$$

$$F(\boldsymbol{x}) = \underbrace{w_d f_d(\boldsymbol{x_d})}_{Depth} + \underbrace{w_r f_r(\boldsymbol{x_r})}_{Reflectance} \tag{2}$$

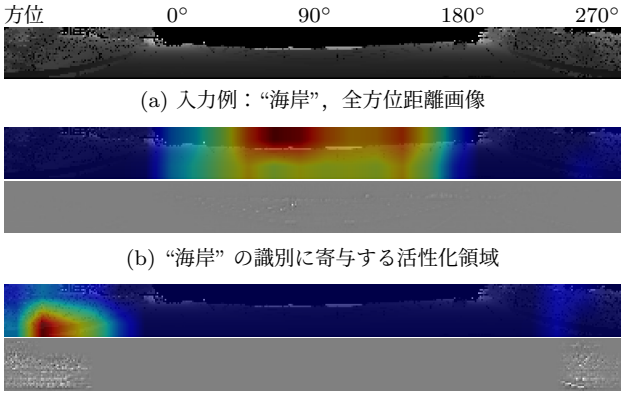
$$(w_d, w_r) = g(r_d(\boldsymbol{x_d}), r_r(\boldsymbol{x_r})) \tag{3}$$

3.4 実装

各種モデルの実装には，深層学習フレームワーク PyTorch を用い，識別評価を行うための計算環境として 1 台の NVIDIA GeForce GTX Titan X を使用した．循環畳込み層は，既存実装のパディングなし畳込み層の前段に循環パディングを新規追加することで実現した．

表 1: 循環畳み込みと RWMP の効果

モデル		循環畳込み	RWMP	正答率 [%]
距離	A			**97.18**
	B		✔	97.11
	C	✔		96.89
	D	✔	✔	96.92
反射強度	A			94.75
	B		✔	95.74
	C	✔		95.45
	D	✔	✔	**95.92**



(a) 入力例：“海岸”，全方位距離画像

(b) “海岸” の識別に寄与する活性化領域

(c) “森林” の識別に寄与する活性化領域

図 3: Grad-CAM による活性化領域（上）と GBP と組み合わせた該当領域における画像特徴の可視化（下）

4. 識別評価実験

4.1 データセットと学習設定

提案手法の性能評価には, Sparse-MPO [1] を用いる．識別対象カテゴリは，“海岸”，“森林”，“屋内駐車場”，“屋外駐車場”，“住宅地”，“市街地” の 6 種類であり，福岡県福岡市の車道 60 箇所を計測した 34200 組の全方位画像から構成される．実験ではまず，計測箇所がオーバラップしないように，ランダムに 10 個のサブセットに分ける．次に，8 セットを学習セット，1 セットをパラメータ調整のための検証セット，残り 1 セットを評価セットとして，10-fold cross validation により評価する．学習アルゴリズムには，Momentum SGD を採用した．学習係数は 0.0001 に固定し，Momentum 係数は 0.9 とした．また，過学習緩和のために，係数を 0.0005 として L2 正則化を適用した．最終層を除く全結合層には Dropout を適用し，学習時に層間結合をランダムに 50%欠落させる．検証セットに対する損失が 10 エポック改善しない場合に，学習を終了する．

4.2 単一種の全方位画像による識別

距離画像と反射強度画像のいずれかを用いた場合の識別評価実験について述べる．表 1 は，VGG モデルに循環畳込み層と RWMP を適用した場合の識別精度の効果を示している．反射強度画像を入力した場合は，循環畳込みと RWMP による効果が見られたが，距離画像を入力とした場合はいずれも正答率の低下を招いた．全体の傾向として距離画像を用いた場合の方が識別精度が高い．また, 表 2 上段に, 距離画像入力と反射強度画像入力それぞれで最も正答率の高い VGG “A”（導入なし），VGG “D”（循環畳込みと RWMP を導入）のカテゴリ毎の正答率を示している．“森林”，“屋外駐車場” に関しては，反射強度画像を入力とした場合の正答率が高く，距離画像と融合することで精度向上が期待できる．

4.3 特徴抽出領域の可視化

循環畳み込みと RWMP による CNN 内部表現への効果を調べるため，Gradient-weighted Class Activation Mapping (Grad-CAM) [10] と Guided Backpropagation (GBP) [12] を適用して特徴抽出領域の可視化を行う．特に，正答率が低下した距離画像入力に焦点を絞る．Grad-CAM は，カテゴリ特有の勾配情報により中間層の特徴マップを重み付けすることで，任意のカテゴリに関して活性化する画像領域を可視化する手法である．本稿では，最終畳込み層の特徴マップを利用した．また，GBP は，特定カテゴリの最終層出力を正勾配のみを利用して逆伝搬し，入力画像を再構成する．CNN が活性化する画像特徴を鮮明に可視化する手法であり，Grad-CAM と組み合わせることができる．図 3 に，“海岸” カテゴリの一例を対象とした可視化結果を示す．入力画像の中央付近（計測車両側面）は，空と海面により点群が取得できていない領域であり，画像両端の陸上側は木々に覆われている．可視化結果から，点群欠落部が “海岸” 特有の特徴として活性化している一方で，画像両端の木々のテクスチャが “森林” の特徴として働いていることが分かる．

図 4 には，評価セットに含まれる全ての距離画像を用いた可視化結果を示している．ここでは, 正解カテゴリに関する活性化領域を Grad-CAM により求め，平均画像を生成している．通常の VGG モデル “A” に比べて，循環畳込みと RWMP を適用した場合（“C”，“D”）は全方位に着目していることが分かる．ただし，“海岸” カテゴリに注目すると，適用後も比較的中央領域が重視される傾向にある．正答率が低下した一要因として，“海岸” のように画像の一部に固有の特徴を有する場合に，受容野を拡大した CNN 内で他カテゴリ特徴の活性化が干渉することが考えられる．

4.4 マルチモーダル全方位画像による識別

距離画像と反射強度画像の両方を用いた場合の識別評価実験について述べる．Softmax average と Adaptive fusion は，第 4.2 節の実験で最も正答率の高い学習済みモデルを利用した．表 2 に，各モデルの正答率を示す．Early fusion を除く全てのモデルにおいて，正答率改善が見られた．特に，Softmax average の平均正答率が最も高く，4 つのカテゴリで最高正答率を示している．Early fusion は，誤差逆伝搬の性質上，モダリティを融合する初段の学習困難性が要因となりうることが報告されており [13]，本実験でも同様のものと考えられる．

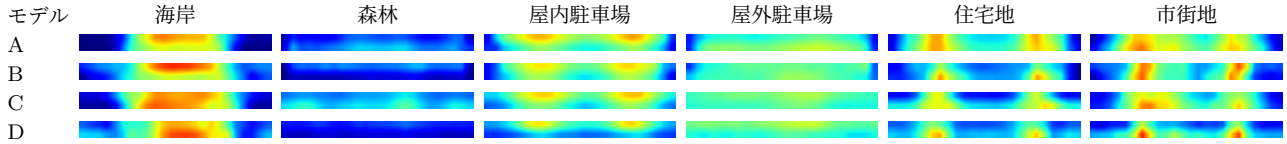


図 4: 評価セットに含まれる全距離画像を用いた活性化領域の平均画像（モデル名は表 1 に対応）

表 2: 各モデルの正答率 [%] （D：距離，R：反射強度，M：距離と反射強度の組み合わせ）

モデル	入力	海岸	森林	屋内駐車場	屋外駐車場	住宅地	市街地	全体
Spin Image [1]	D	65.60	86.30	81.84	86.26	82.95	64.31	79.23
LBP [1]	D	84.25	94.93	96.41	86.86	94.58	92.71	92.00
VGG “A”	D	**92.73**	97.26	**99.94**	94.23	**98.35**	**99.20**	**97.18**
VGG “D”	R	91.83	**98.20**	91.45	**95.16**	97.99	98.27	95.92
Early fusion	M	93.37	98.09	99.44	94.71	98.21	97.15	97.02
Late fusion	M	93.49	97.57	99.25	94.23	98.39	98.88	97.19
Softmax average	M	94.27	**98.38**	99.58	**94.91**	99.12	**99.56**	**97.87**
Adaptive fusion	M	**94.59**	98.20	**99.77**	94.85	**99.19**	99.37	97.62

5. まとめと今後の展望

本稿では，全方位型 3D LiDAR から得られる距離画像と反射強度画像を用いた一般屋外環境識別手法を提案した．循環畳込み層および RWMP 層の導入により，反射強度画像のみを用いた実験では，通常の CNN に比べて正答率が向上した．また，CNN の活性化領域を可視化したことで，循環畳込み層および RWMP 層の効果を確認し，本データセットにおける問題点を考察した．距離画像と反射強度画像を組み合わせた実験では，単一種の画像を用いた場合よりも高精度な識別結果を得た．本稿では VGG-11 モデルをベースラインとして検証を行ったが，今後は他のアーキテクチャを利用した実験を行う予定である．

謝辞

本研究は文部科学省科学研究費補助金基盤研究（A）（課題番号 26249029）の支援を受けた．

参 考 文 献

[1] H. Jung, Y. Oto, O. M. Mozos, Y. Iwashita and R. Kurazume: “Multi-modal Panoramic 3D Outdoor Datasets for Place Categorization”, *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp.4545–4550, 2016.

[2] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba and A. Oliva: “Learning deep features for scene recognition using places database”, *Advances in Neural Information Processing Systems (NIPS)*, pp.487–495, 2014.

[3] S. Song, S. P. Lichtenberg and J. Xiao: “SUN RGB-D: A RGB-D scene understanding benchmark suite”, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.567–576, 2015

[4] B. Shi, S. Bai, Z. Zhou and X. Bai: “DeepPano: Deep Panoramic Representation for 3-D Shape Recognition”, *IEEE Signal Processing Letters*, vol.22, no.12, pp.2339–2343, 2015.

[5] O. Mees, A. Eitel and W. Burgard: “Choosing Smartly: Adaptive Multimodal Fusion for Object Detection in Changing Environments”, *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp.151–156, 2016.

[6] K. Simonyan and A. Zisserman: “Very Deep Convolutional Networks for Large-Scale Image Recognition”, *Proc. Int. Conf. on Learning Representations (ICLR)*, 2015

[7] D. Maturana and D. Scherer: “VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition”, *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp.922–928, 2015

[8] S. Gupta, R. Girshick, P. Arbeláez and J. Malik: “Learning Rich Features from RGB-D Images for Object Detection and Segmentation”, *Proc. European Conf. on Computer Vision (ECCV)*, pp.345–360, 2014

[9] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller and W. Burgard: “Multimodal Deep Learning for Robust RGB-D Object Recognition”, *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pp.681–687, 2015.

[10] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra: “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”, *ArXiv e-prints*, 1610.02391, 2016.

[11] S. Ioffe and C. Szegedy: “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”, *Proc. 32nd Int. Conf. on Machine Learning (ICML)*, pp.448–456, 2015.

[12] J. T. Springenberg, A. Dosovitskiy, T. Brox and M. Riedmiller: “Striving for Simplicity: The All Convolutional Net”, *Int. Conf. on Learning Representations (ICLR)*, 2015.

[13] J. Long, E. Shelhamer and T. Darrell: “Fully Convolutional Networks for Semantic Segmentation”, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.3431–3440, 2015