全方位画像からの推定深度情報を用いた深層学習による空間の評価予測

Deep Learning Model for Predicting Space Evaluation by Estimating Depth Information Using Omnidirectional Images

○衣川　雛[*1]，瀧澤　重志[*2]
Hina Kinugawa[*1] and Atsushi Takizawa[*2]

*1 大阪市立大学大学院生活科学研究科　前期博士課程
Graduate Student, Graduate School of Human Life Science, Osaka City University.
*2 大阪市立大学大学院生活科学研究科　教授　博士(工学)
Professor, Graduate School of Human Life Science, Osaka City University, Ph.D.

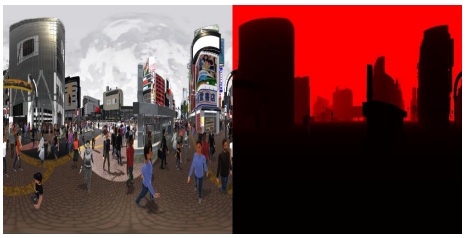
キーワード：全方位画像; 深度画像; Unity; Google ストリートビュー; pix2pix; RankNet
Keywords: Omnidirectional image; depth image; Unity; Google Street View; pix2pix; RankNet

1. はじめに

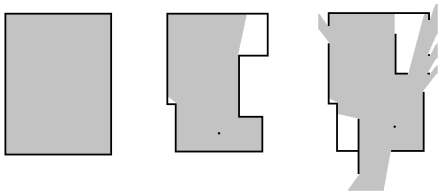
都市計画やランドスケープの分野では，景観の印象や環境工学的な評価を行うために，画像を用いた分析が広く行われている．このような研究を行うには，地図に紐づいた網羅的な画像データベースが必要になるが，Google Street View (GSV)のような大規模な画像データが整い，このような分析が意味を成すようになってきている．一般に，画像を使用して空間を評価する場合，画像から意味のある計算可能な指標を抽出する必要がある．しかし，特に空間の印象を考えた場合，それに影響を与える画像の特徴は，複雑であったりあいまいであったりと，必ずしも明示的に定義することができない性質を有していると考えられる．

近年になって，画像から特徴量を自動的に抽出し，分類，回帰等に用いる，深層畳み込みニューラルネットワーク（Deep Convolutional Neural Network: DCNN）の技術が急速に発展しており，現在の AI ブームの基盤技術となっている．大量の GSV の画像と DCNN を使って，都市の景観評価をモデル化した例として，Liu ら[1]，Law ら[2] 等の研究がある．彼らの研究では，通常の画角の RGB 画像を使っている．RGB 画像は色やテクスチャ的な特徴を捉えるのに適していると考えられる．一方，空間は見通しや広さといった幾何的な特徴も重要である．こうした空間の特徴は，従来からスペースシンタックスと呼ばれる空間分析方法によって研究がなされてきたが，現在の機械学習の技術では，直接幾何的な特徴を捉えることは難しい．

Takizawa ら[3] は，ゲームエンジンの Unity で構築された CG の都市空間内で，全方位の RGB 画像と深度画像を多数撮影し（図-1），これらの画像を入力データとして，DCNN により別途行った仮想都市景観の評価実験結果を推定する評価モデルを作成した．その結果，RGB 画像に深度画像を追加することで，分類精度の向上や人の判断に近い CNN が学習できることを明らかにした．ちなみにこの全方位の深度画像はいわゆる Isovist（図-2）[4] の 3 次元バージョンであり，これまで幾何学的なアプローチで分析されてきたが，DCNN を用いることにより，特徴量を明示することなく画像ベースで分析している点で，新しい分析方法の提案になっている．



図－１　全方位の RGB 画像（左）とその深度画像（右）
深度画像は色が黒いほど近く，赤くなるほど遠景となる



図－２　2 次元の Isovist

しかし我々が検証を行いたい現実の都市空間の深度画像を得ることは容易ではない．例えば，3 次元レーザースキャナーやフォトグラメトリを使用しても，低コストで網羅的に都市スケールでの深度画像を得ることは依然として困難である．だが CG では正確な深度画像が生成できる．

以上の背景から本研究では，CG で作成した全方位の RGB 画像と深度画像のペアを，深層学習による汎用的な画像変換手法である pix2pix[5] を用いて学習し，全方位の RGB 空間画像から深度画像を生成する方法を提案する．そして，条件を変えて学習したモデルを，GSV の画像に適用して深度画像を生成し，その妥当性を目視で評価する．さらに，GSV の景観画像を複数の被験者で評価する実験を行う．そ

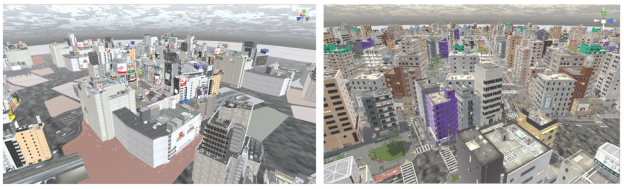
日本建築学会情報システム技術委員会
第42回情報・システム・利用・技術シンポジウム論文集，176-179，2019年12月，東京
Proceedings of the 42nd Symposium on Computer Technology of Information, Systems and Applications，AIJ，176-179，Dec.，2019，Tokyo

の評価値を，CNN のランク学習によって深度画像を含めた/含めない GSV の画像から推定するモデルを構築する．そして，深度画像の有無により，この景観評価モデルの汎化性能がどの程度変化するのかを確認し，実空間における深度画像の有効性を確認する．

2. 提案方法 1：深度画像の生成

2.1. 3D の都市空間モデルの構築

既往研究と同様にゲームエンジンの Unity を用い，対象となる都市の 3D モデルを用意する．3D モデルとして，CG マーケットで販売されている 3D 都市モデルのデータセット 7)の中から，リアリティが高く実際の日本の都市空間に忠実なものを選んで使用した．本研究では，渋谷を模した渋谷モデル（図-3a）と，日本の地方の市街地を模した地方都市モデル（図-3b）の 2 つのモデルを使用した．また，GSV 画像に学習したモデルを適用することを踏まえ，CG の 3D モデルを実際の街路空間の状況に近づけるために，街路上に人や車のモデルを設置した．オリジナルの 3D モデルの空間の範囲は，現実の空間に近い距離画像を取得するほどには広くないため，同じモデルをコピー・拡張し，遠方のビル等を疑似的に表現した．なお，教師データとして使用する深度画像は，奥行きの空間情報のみを取得するために，人や車などの空間にとって障害物となるオブジェクトを除いて生成させている．また，実際の風景写真は，同じ場所の天気，季節，時間帯によって大きく変化するため，Unity の天気アセットを使用して，太陽の高度や空などの環境条件を変更し，青空と曇り空のモデルを設定した．



(a) 渋谷モデル　　(b) 地方都市モデル

図－３ 使用した敷地モデル

2.2. 全方位画像の撮影

用意した 2 つの都市モデルのそれぞれで，カメラの高さを 2.05m に設定した．これは日本の GSV の車載カメラの高さである．次に，図-4 のように 3D モデルの平面画像を GIS に入力し，路上を中心に撮影スポットを各モデルで 550 点ずつランダムに設定した．それらの観測地点の座標を Unity に読み込み，各地点で全方位の RGB 画像と深度画像を撮影する．全方位画像の撮影には，Unity のカメラアセットである Spherical Image Cam（現在は販売中止）を利用した．同一地点でカメラを 20 度ずつ回転させて 18 枚の全方位画像を撮影することで，データ数を増やしている．

図－４ 地方都市モデルにおける撮影地点

2.3. 学習用のデータセットの作成

pix2pix を使用して，全方向 RGB 画像から深度画像を生成するためのモデルを学習させる．pix2pix は深層学習に基づく画像生成アルゴリズムの一種で，ペアの画像から画像間の関係を学習することで，1 枚の画像からその関係を考慮して対となる画像を生成する技術である．

深度画像の生成結果を図-5 に示す．例えば，図-5a に示した RGB 画像と図-5b の深度画像のような 2 枚の画像のペアの集合をもとに，画像の関係性を学習させたモデルに対して，図-5a を入力すると，図 5c のような推定された深度画像が生成される．学習のためのデータセットは，二つの空間モデルと，それらを組み合わせた混合空間モデルの 3 通りを考える．同様に，空のモデルを，青空，曇天，それらを組み合わせた混合空モデルの 3 通りを考える．最終的に，表-1 に示した 9 通りの空間モデルを準備し，それぞれで全方位画像を撮影して学習用，検証用の画像データセットを構築した．学習用と検証用のデータの比率はすべてのモデルで 3:1 とした．建物内部が映り込んでしまった画像を省き，最終的に M1a，M1b では合計 10044 枚の画像を，M2a，M2b では合計 3925 枚の画像を使用した．

(a)入力 RGB 画像　(b)入力深度画像　(c)推定深度画像

図－５ 深度画像の生成結果

表－１ 学習に用いたモデルの組み合わせ

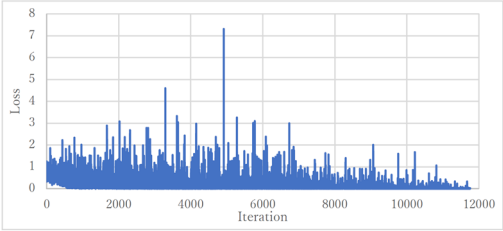
空間モデル	空モデル		
	a.青空	b.曇天	c.混合
1. 渋谷	M1a	M1b	M1c
2. 地方都市	M2a	M2b	M2c
3. 混合	M3a	M3b	M3c

2.4. 学習結果と検証

2.4.1 pix2pix の学習

pix2pix は PyTorch[8] で実装されたもの [9] を用い，学習のエポック数を 400，その他はデフォルトの設定として学習

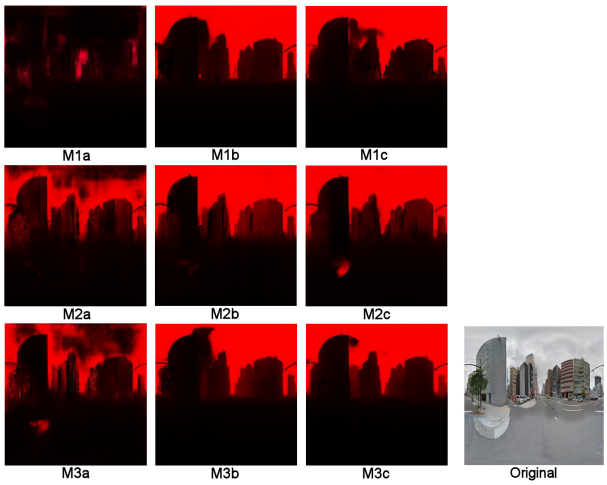
を行った．学習時に得られる損失関数のグラフから精度を評価する．pix2pix では，画像生成を行う生成器と，その画像が本物か偽物かを判定する判定器から構成されている．今回は，画像の真偽を評価する判定器の損失関数の収束度合いから精度を確認した．図-6 は，M2b を 400 エポックまで学習させた際の，損失関数（D_fake）の収束例である．若干の揺らぎが見られるものの，学習の進展により損失関数の値が減少していくことが見てとれる．



図－６　損失関数の例（D_fake）

2.4.2 GSV による学習モデルの検証

GSV の全方位画像を入力し，生成された深度画像がどの程度現実に即しているかを目視により評価する．結果の例を図-7 に示す．表-1 の 9 つのモデルの中で，最も結果が良かったのは M2b モデルであった．そのほかのモデルでは，特に空の部分の距離が不正確になってしまっている．



図－７　異なるモデルによる推定深度画像の例

3. 提案方法 2：推定深度画像とランク学習による GSV 画像の評価予測

3.1. 印象評価実験

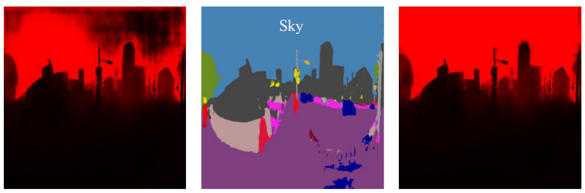
3.3 に示すランク学習のデータとするために，GSV の画像を Oculus Rift に投影し，印象評価実験を行う．GSV の画像は大阪府市内の住宅街と繁華街から，それぞれ 50 枚ずつサンプリングし，合計 100 枚の画像を用いた．収集した GSV の景観を，｛よい ＝4，やや良い ＝3，やや悪い ＝2，悪い ＝1｝の 4 段階で評価する．10 人の建築系の学生を被験者とし，疲労や飽きの問題を考慮して各被験者が 50 枚の画像を評価することとした．各画像の評価値の平均値をその画像の評価値としてランク学習に用いる．評価値が高い場所，平均的な場所，低い場所の画像例を図-8 に示す．

(a) 高 (3.4)　(b) 中 (2.2)　(c) 低 (1.0)

図－８　印象評価実験による評価値が異なる画像の例

3.2. GSV 画像の RGBD 画像の作成

前述した深度画像作成モデルにより，被験者実験で用いた 100 枚の GSV の各画像について，深度の推定画像を作成した．この時，目視で最も結果が良かった Japan Blocks の曇り空モデルを用いて推定を行った．しかし，図-9a のように，空部分の誤推定が目立ったため，Semantic Segmentation (SS) により空部分を抽出し(図-9b)，図-9a の該当ピクセルの値を最大値（赤色）とするフィルタリング処理を行った(図-9c)．SS のモデルには，DeepLab v3[10] の xception71_dpc_cityscapes_trainval を用いた．そして，元の GSV の RGB 画像を，1 チャンネルあたり 8 ビットの合計 4 チャンネルの RGBA フォーマットに変換し，それを 256*256 ピクセルのサイズにリサイズし，同サイズの深度画像の R 成分の値を，その Alpha チャンネルに格納することで，RGBD の 4 チャンネルの画像を作成した．次に，カメラの撮影角度によるバイアスを減らすために，各画像を同一地点で 36 度ずつ回転させ 10 枚の画像に増やす．このようにして合計 1000 枚の GSV の RGBD 画像を作成した．



(a) 推定深度画像　(b)SS 画像　(c)フィルタリング後の推定深度画像

図－９　深度画像のフィルタリング例

3.3. DCNN によるランク学習モデル

先に求めた景観画像の評価値を画像から推定するモデルを作成するために，CNN を用いたランク学習モデルを適用する．ここでは，データのペアの評価値の大小を相対的に比較する RankNet[11] の考え方でランク学習を行う．画像が最大で 4 チャンネルの入力データなので，一般的な画像分類の CNN をベースとして，その画像入力層のチャン

ネル数を 3 から 4 チャンネルに変更し，さらに出力が実数のスカラー値なので，最終層のシグモイド関数の全結合層を，単純な線形結合でスカラー値を返す層に変更した．基本となる CNN に，1000 クラスの ImageNet で事前学習が行われた Resnet-152[12] を用い，その全パラメータをファインチューニングする形で学習を行う．各トレーニング画像に対して，その他の画像を 1 枚ランダムサンプリングしてペアを作ってロスを求めることを，すべてのトレーニング画像で行うことで，1 エポックとした．

撮影地点が 100 地点と深層学習を行うには少ないので，各撮影地点についてカメラの角度が 0 度の画像をテストデータとし，残り 900 枚の画像について，75 地点の画像を学習データ，25 地点の画像を検証データとして画像をランダムに分割した．結果，トレーニングデータ 675 枚，Validation データ 225 枚，テストデータ 100 枚の構成で学習と検証を行った．深層学習のフレームワークには PyTorch を用いた．学習のハイパーパラメータを以下に示す．エポック数=500，最適化手法=Momentum SGD，モーメント=0.9，学習率=0.001，weight decay＝0.00001，ミニバッチサイズ=338．なお，最終的なモデルの評価は，各データセットのすべての画像の推定された評価値の順位に関するスピアマンの順序相関係数（SRCC）で行った．

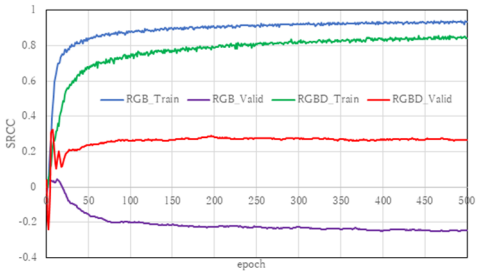
3.4. 学習結果と検証

図-10 に，RGB と RGBD 画像をそれぞれ用いた際の，学習時の学習データと検証データでの SRCC の推移を示す．この図から，学習データについては RGB 画像のほうが精度が高いが，検証データでは，RGB 画像で学習したモデルは，ほとんど予測ができていないどころか，反対の評価を行う誤った方向にモデルが学習されていくことがわかる．一方，RGBD データでは SRCC が 0.3 程度と必ずしも高くは無いが，検証データでもある程度，評価値の予測ができていることがわかる．検証データで最も SRCC が高かったのは RGB データで 13 エポック後，RGBD データでは初期の不安定な学習時を除くと 196 エポック後であった．これらのエポックでのモデルをベストモデルとして，それらをテストデータで評価した際の SRCC を表-2 に示す．テストデータは学習データと同一撮影地点の異なる角度で撮影された画像を含むので，学習データで使った場所と検証データで使った場所に分けて SRCC を求めた．結果は，前述した図-1 の結果を支持するものであり，RGB 画像で学習させたモデルの汎化性能がほとんどない一方で，RGBD 画像で学習させたモデルはその SRCC が 0.41 であり，人の嗜好とやや相関がある評価モデルが作成できたといえる．

以上の結果から，深度画像は現実の空間の写真を利用した空間評価において有用であると結論付けられよう．

4. おわりに

pix2pix を用いて，全方位画像からその深度画像を生成する手法を提案し，CG を学習データとして用いて学習を行い，GSV の画像で検証を行ったところ，学習データによっては，違和感のない深度画像を生成できることが分かった．さらに，生成された深度画像を景観評価のランク学習のデータとして用いることで，学習モデルの汎化性能を大きく向上させることができた．したがって，深度画像が画像を用いた景観評価などに不可欠な情報になりうることが示された．今後は，生成された深度画像の深度の詳細な評価と精度の向上が必要である．



図－１０ 学習データと検証データの SRCC の推移

表－２ テストデータの各モデルによる SRCC (p 値)

画像データ	学習データ	検証データ
RGB	0.743 (p<0.001)	0.045 (p=0.830)
RGBD	0.742 (p<0.001)	0.411 (p=0.041)

[謝辞]
本研究の一部は科研費基盤(A)(B)の補助を受けています．

[参考文献]

1) L. Liu et al., (2017) A machine learning-based method for the large-scale evaluation of the urban environment, Computers, Environment and Urban Systems, pp. 113-125
2) S. Law et al., (2017) An application of convolutional neural network in street image classification, GeoAI'17
3) A. Takizawa et al, (2017) 3D Spatial Analysis Method with First-Person Viewpoint by Deep Convolutional Neural Network with Omnidirectional RGB and Depth Images, eCAADe, pp. 693-702
4) M. Benedikt, (1979) To take hold of space: isovists and isovist fields, Environment and Planning B, 6, pp. 47-65
5) P. Isola et al, (2017) Image-to-Image Translation with Conditional Adversarial Networks, CVPR
6) G. Ros et al., (2016) The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3234-3243
7) NoneCG, https://www.nonecg.com/3D-products/tokyo-shibuya/（2019 年 6 月 17 日閲覧）
8) Pytorch, https://pytorch.org（2019 年 6 月 17 日閲覧）
9) Junyanz, pytorch-CycleGAN-andpix2pix, https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix（2019 年 6 月 17 日閲覧）
10) L. C. Chen et al., (2018) Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, ECCV
11) C. Burges et al., (2005) Learning to Rank Using Gradient Descent, ICML '05, pp.89-96
12) K. He et al., (2016) Deep Residual Learning for Image Recognition, CVPR