マルチモーダル深層学習を用いた街並み画像に対する人間の振る舞い予測
- 注視傾向予測及び結果を付与した多次元データによる訪問意欲予測を対象に -

環境都市専攻 建築都市デザインコース 6143190004-0 大野 耕太郎
（指導教員 山田 悟史）

1. はじめに

深層学習を基盤とする人工知能（AI）技術が幅広い分野で研究されている。画像を例にすれば深層学習の特徴は，画像の意味を表現する特徴量と因果関係自体を AI が学習することにある。この特徴を活かし，様々な分野で既存の結果を上回る成果が報告されており，筆者らも既往研究で感性に対する適用可能性を提示している。これら多くは非構造な画素値を用いて分類・推定を行っているが，画像を用いることには限界があり，また合理的でない。このような視点から提案されている学習モデルに「マルチモーダル深層学習」がある。

上記の先端的な深層学習モデルの建築・都市分野に対する適用可能性は高い。それは，人間も「視覚」「聴覚」「嗅覚」といった複数の情報や，経験から形成される対象に対する認知も含めて高次元的に意思決定をしているからである。このような可能性の中で本研究が着目しているのは「デザインに対する人間の振る舞い予測」である。視覚を通して抱く印象・感性はデザインの重要な目的関数であるが，視覚を通して認識する空間と印象・感性という複雑な関係は，学術的にも実務的にも未解明な部分が残る。

以上より，本研究では街並みの画像を対象に，訪問意欲を推定する AI の作成と注視点情報という人間の生理反応を情報付与に用いたマルチモーダル深層学習による AI の作成を検証する。

2. 研究概要

2.1 学習データセット

データセットとして研究に用いる街並み画像は，大都市や観光地から選定した表 1 に示す 21 街路から作成した 2100 枚である※1。

2.2 被験者実験による注視点とスケッチの取得

表 1 学習データセットに用いる街並み一覧

Class	Country	City	回答数	訪問意欲度合	正規化ラベル
1	America	New York	25/46	0.543478	0.570777
2	Australia	Melbourne	15/46	0.326086	0.276414
3	Canada	St John's	18/45	0.4	0.376497
4	Czech	Prague	38/45	0.844444	0.978306
5	England	London	37/45	0.822222	0.948216
6	Hong Kong	Mong Kok	15/44	0.340909	0.296484
7	Italy	Firenze	36/42	0.857142	0.995501
8	Japan	Kyoto	37/43	0.860465	1
9	Japan	Osaka	7/43	0.166666	0.060547
10	Japan	Tokyo	5/41	0.121951	0
11	Mexico	Guanajuato	30/41	0.731707	0.825652
12	Peru	Cusco	28/41	0.682926	0.759600
13	Portugal	Porto	31/42	0.738095	0.834302
14	Russia	Moscow	31/42	0.738095	0.834302
15	Scotland	Edinburgh	35/42	0.833333	0.963261
16	South Africa	Cape Town	25/43	0.581395	0.622119
17	Korea	Seoul	17/44	0.386363	0.358033
18	Spain	Barcelona	26/44	0.590909	0.635002
19	Taiwan	Chiufen	22/44	0.5	0.511904
20	Thailand	Bangkok	13/44	0.295454	0.234935
21	Arab	Dubai	28/44	0.636363	0.696550

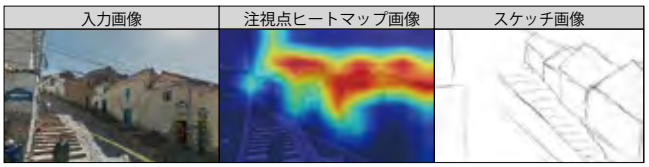


図 1 注視点ヒートマップ画像とスケッチ画像一例

被験者実験を行い，マルチモーダル学習に用いる注視点情報の取得及び街並み画像を見た際の集団の街路に対する集団の訪問意欲の度合を算出した（表 1)。また同時にスケッチ画像の取得も行う有効回答として各 910 枚分のデータセットを得た（図 1)※2。

2.3 深層学習モデルの説明

図 2 に本研究で用いる深層学習モデルのネットワーク図を示す。深層学習において高い精度を誇るモデルとして知られる「VGG」を参考に縦 200 ピクセル，横 338 ピクセルの RGB 画像から畳み込みニューラルネットワーク（CNN）を介し，0（訪問したくない）～ 1（訪問したい）までの範囲で街並み画像に対する訪問意欲の度合いを推定する回帰学習

Human Behavior Prediction for Cityscape Images Using Multimodal Deep Learning

- For prediction of gazing tendency and prediction of willingness to visit using multi-dimensional data with results-

ONO Kotaro

型のモデルを作成した。

2.4　街並みの印象・感性評価推定 AI の作成

筆者らによる事前検証[1]により，VGG ベース深層学習モデルでの「街路名分類」「個人の訪問意欲の分類・推定」において高い精度での分類・推定が可能であることを確認した。そこで，追加検証として集団の街並みに対する訪問意欲の度合い推定 AI の作成の可否を検証する。学習データは左右反転による水増しを行った街並み画像 4200 枚である※3。1000epoch の学習を行い，正解ラベル（正規化された訪問意欲の度合い）と AI の推定値との平均二乗誤差 (MSE) を損失関数として使用する。また，作成したモデルの精度検証には，MSE が最も小さかったモデルを用いる（以下，最良モデルと呼称）。

2.5 注視点情報を付与したマルチモーダル学習

通常の画像を用いた集団の訪問意欲の度合いの推定が可能であることを確認したのちに，被験者実験によって得られた注視点情報を画像に追加情報として付与したマルチモーダル深層学習を検証する。情報の付加の手法として，「画像への注視点情報マスキング処理」「RGB 画像に注視点情報を加えた次元追加」の２パターンを検証する。図３に情報付与のイメージ図を示す。「画像への注視点情報マスキング処理」では，注視点画像を元の街並み画像にマスクイメージとして追加することで，人間が注視している領域を AI が特徴量として学習する。また，衣川らの研究[2]では Google ストリートビューの RGB 画像に対し，深度情報 (D) をアルファチャンネルとする RGBD4 次元情報を用いた都市の景観評価を分類する AI を作成し，３次元情報と比較した際に次元追加モデルでは一定の効果があることを示している。そこで本研究では，深度情報よりも人間の感性評価と直接的な関係性を持つと考えられる注視点情報を街並み画像に次元追加した精度向上を試みた。

2.5.1 被験者実験画像を用いたマルチモーダル学習

まず注視点情報が得られた 910 枚の街並み画像を左右反転し，1820 枚の画像に対して集団の訪問意欲の度合いを AI 学習時のラベルとして付与する。通常画像を用いたモデルを加えた３パターン作成し，1000epoch 学習時の最良モデルを用いた比較・考察を行うことで本手法の有効性を検証する※4。

2.5.2 注視点傾向を予測する AI

注視点情報の取得には専用のデバイスを用いてキャリブレーションによる視線推定を行う必要があり，大量のデータセットを作成するのは困難である。また，街並みの画像というデザイン要素を観た際の人間の注視傾向という生理反応の予測にはいまだ不明瞭な点も多く存在する。そこで本章では街並み画像から，その画像に対する注視傾向を予測する AI の作成を検証する。注視点画像の生成には画像変換を行う深層学習アルゴリズムとして広く用いられる pix2pix[3] を使用する※5。

2.5.3 注視点予測結果を用いたマルチモーダル学習

最後に，被験者実験には用いられていない画像を含む，左右反転で水増しを行った街並み画像 4200 枚に対し，pix2pix によって作成された注視傾向予測画像を情報付与したマルチモーダル学習を行う。デバイスによる詳細な注視点の計測を行わずに予測された注視点データにおいても，本研究が提唱するマルチモーダル学習が有効かを確認する。1820 枚での学習時と同様に，1000epoch 学習を行いマスキング・次元追加の２パターンと，4200 枚通常画像モデルの最良モデルを比較する。

３．検証結果

3.1　集団の訪問意欲の度合い推定 (4200 枚)

図３に通常画像学習モデルでの 1000epoch 学習を行った際の学習の推移と，最も誤差が小さかったモデル（以下最良モデルと呼称）でのプロット図を示す。学習の推移では，300epoch 付近で収束がみられ，最良モデルの平均二乗誤差 (MSE) は 0.0197 だった。Pearson の積率相関係数 (PRCC) を用いて被験者 (Human) と再正規化を行った AI の推定値を比較した結果，相関係数 0.8890 となり，画像に対する推定モデルの構築に成功した。一方で，誤差の絶対値の平均である平均絶対誤差 (MAE) は 0.0791 となり画像単体では平均± 8% 程度の誤差があった。

3.2 マルチモーダル学習 (1820 枚)

図５～図７に 1820 枚学習モデルでの情報付与パターン別の，学習の推移と最良モデルでのプロット図を示す。MSE の結果では，次元追加モデルでは 0.0247 と通常画像学習モデルよりも 2 割程度の減

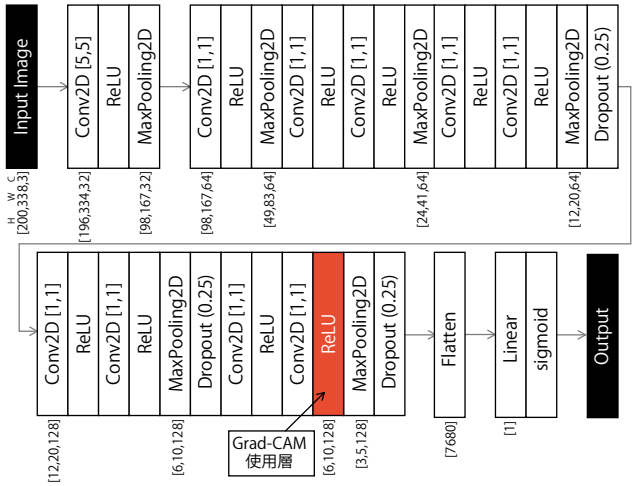


図 2　ネットワーク図

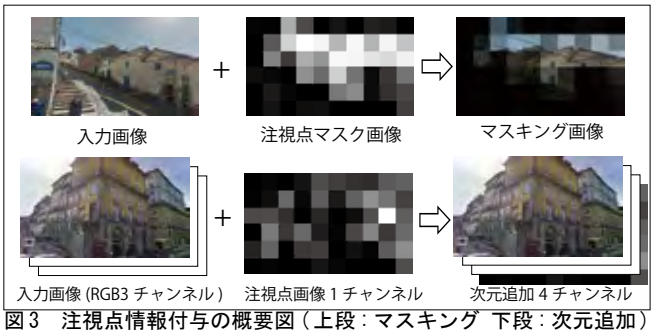
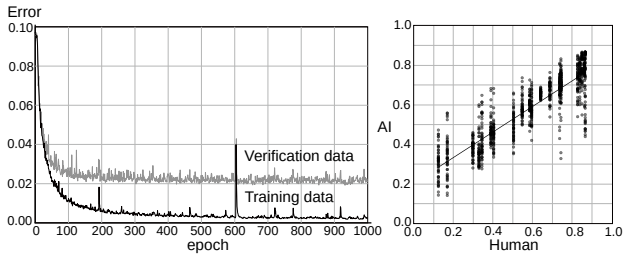


図 3　注視点情報付与の概要図（上段：マスキング　下段：次元追加）

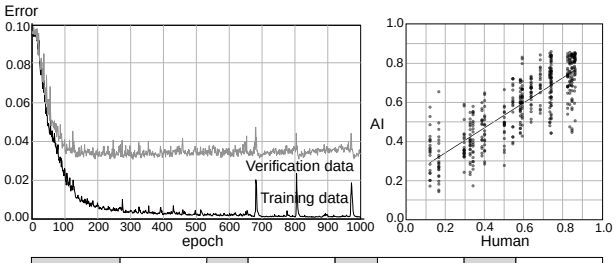


Best epoch	932	MSE	0.0197	MAE	0.0791	PRCC	0.8861

図 4　4200 枚通常画像学習モデルでの学習推移（左）と散布図（右）

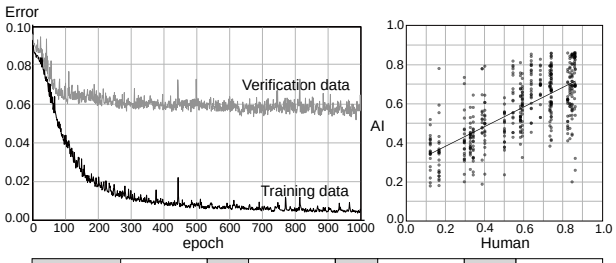
少が見られた。また PRCC でも 0.8230 から 0.8661 と 0.04 の向上が，MAE では 0.0972 から 0.0860 と 12%近い減少が確認できた。次元追加モデルには一定の効果があるといえる。その一方でマスキングモデルでの数値は，すべてにおいて通常画像学習モデルの結果を下回った。

また，3 パターンにおける AI の注目領域を，可視化手法の一種である Grad-CAM[4] を用いて分析を行った。可視化には図 8 のように正規化された 6*10 サイズのデータを用いる。図 9 に Grad-CAM の結果一例を，表 2 に 100epoch ごとの Grad-CAM 結果と注視点情報との PRCC 平均を示す。PRCC はマスキングモデルが 0.5 程度と高い値をとった。次元追加モデルでは，通常画像と比較して PRCC の値が高く



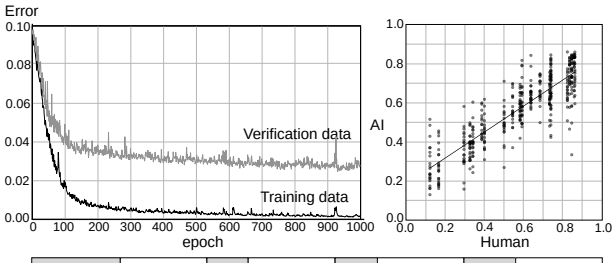
Best epoch	308	MSE	0.0300	MAE	0.0972	PRCC	0.8230

図 5　1820 枚通常学習モデルでの学習推移（左）と散布図（右）



Best epoch	630	MSE	0.0512	MAE	0.1209	PRCC	0.6807

図 6　1820 枚マスキングモデルでの学習推移（左）と散布図（右）



Best epoch	931	MSE	0.0247	MAE	0.0860	PRCC	0.8661

図 7　1820 枚次元追加モデルでの学習推移（左）と散布図（右）

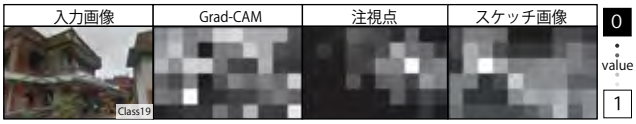


図 8　Grad-CAM の出力結果に合わせた画像処理の一例

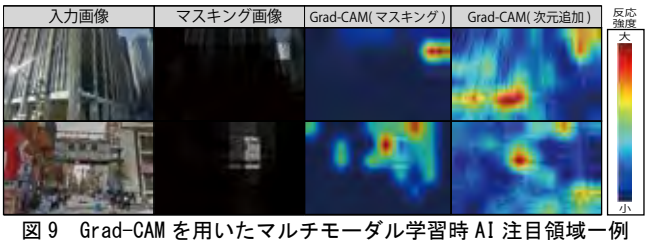


図 9　Grad-CAM を用いたマルチモーダル学習時 AI 注目領域一例

表 2　100epoch ごとの Grad-CAM 結果と注視点情報との PRCC 平均

epoch	100	200	300	400	500	600	700	800	900	1000
通常画像	0.0225	-0.0179	-0.0119	-0.0240	-0.0073	0.0116	-0.0107	-0.0173	-0.0055	-0.0139
マスキング	0.6374	0.6414	0.5975	0.5893	0.5763	0.5594	0.5347	0.5558	0.5301	0.5476
次元追加	0.2652	0.2617	0.1570	0.1279	0.1163	0.1504	0.1851	0.1234	0.0686	0.0989

なっており，可視化結果を見ると通常画像学習モデルとマスキングモデルを合わせたような領域に着目しており，広く特徴量の伝達を行った結果，学習の向上につながってた可能性が考えられる。

3.3 pix2pix による注視傾向予測

図 10 に pix2pix によって生成された注視傾向予測画像の一例を示す。実際の注視点画像と比較すると，近しい位置に推定をすることができている。

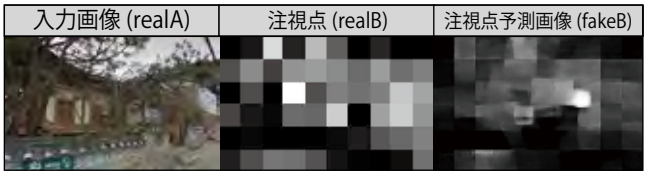
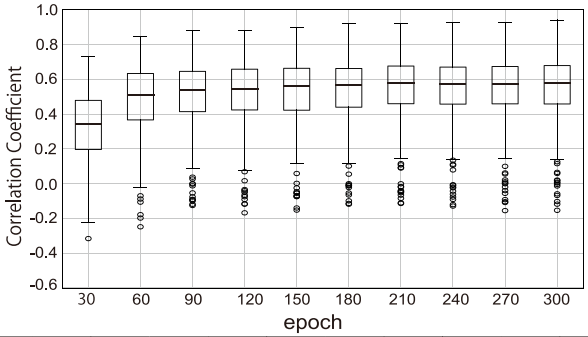


図 10 pix2pix による注視点予測画像生成結果一例



epoch	30	60	90	120	150	180	210	240	270	300
Score	0.3262	0.4771	0.5118	0.5228	0.5298	0.5342	0.5469	0.5441	0.5445	0.5517

検証用画像に対する相関係数の平均値

図 11　pix2pix30epoch ごとの注視点画像との相関係数箱ひげ図

図 11 に 300epoch までの 30epoch ごとの注視傾向予測画像と実際の画像との PRCC の箱ひげ図を示す。推移を見ると，PRCC 平均値は 0.50 付近で安定し，300epoch で 0.5517 となった。

3.4 マルチモーダル学習（4200 枚）

図 12，図 13 に pix2pix 最良モデルで 4200 枚の街並み画像に注視点情報を付与し，マルチモーダル学習を行った際の学習の推移と最良モデルでのプロット図を示す。1820 枚画像モデルと同様に次元追加モデルでは PRCC 結果 0.9024 など，1820 精度の向上が確認された。一方，1820 枚画像学習モデルでの情報付与時の学習効果よりは低い結果となった。pix2pix の予測精度に依存していると考えられる。

4．まとめ

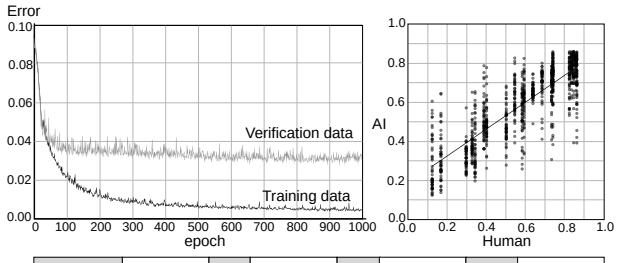
３パターンの検証の結果から，注視点情報を次元追加したマルチモーダル深層学習で一定の成果を得た。情報付与の種類や方法を追加検証することで，さらなる精度の上昇が可能かを今後の課題とする。

注釈

※ 1) 画像が単一の建築物・地面・空に占められることのないように配慮しながら Google Earth のストリートビューを用いて街路の範囲内からランダムに作成した。枚数は既往事例と予備実験を参考に 1 街路 100 枚とした。

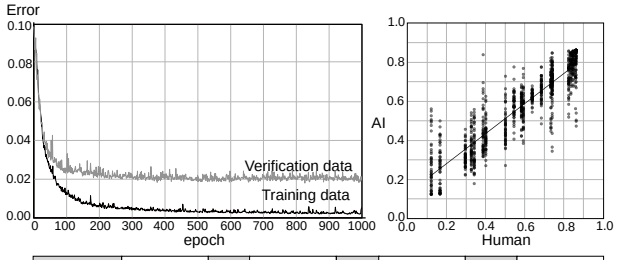
※ 2) 被験者実験は建築系学生 100 名（男性 69 名，女性 31 名）を対象に実施した。画像をモニター越しに 30 秒間見てもらい「訪問したい」あるいは「訪問したくない」の２択で回答してもらった。この際モニターにマーカーを設置し，注視点計測デバイス用いて注視点の取得も行う。その後，被験者は街並み画像を隠した状態で２分間の記憶に基づくスケッチを行う。被験者１名あたり，10 枚の街並み画像に対して回答を行った。

※ 3) 学習用画像 75%(3150 枚），検証用画像 25%（850 枚）に分割し，反転前と反転後の画像が学習時に混在しないようにしている。被験者実験で得られた集団の訪問意欲の度合いを正規化することで 0 ～ 1 までの値に直し，学習時のラベルとして付与した（表 1)。また，同一のクラス（街路）である街並み画像すべてに同一ラベルを付けた。



Best epoch	906	MSE	0.0283	MAE	0.0899	PRCC	0.8399

図 12　4200 枚マスキングモデルでの学習推移（左）と散布図（右）



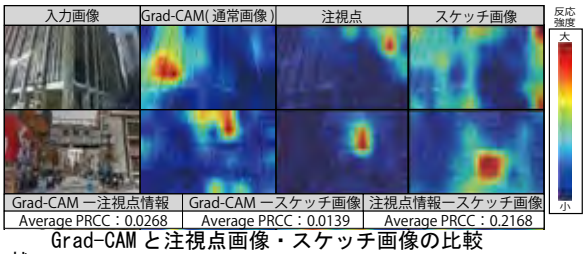
Best epoch	827	MSE	0.0180	MAE	0.0682	PRCC	0.9024

図 13　4200 枚次元追加モデルでの学習推移（左）と散布図（右）

※ 4) 学習時には画像のランダムシャッフルを行い，学習用画像 75%，検証用画像 25% に分割した。また，通常画像モデル，マスキングモデル，次元追加モデルともに同じ条件で学習を行うためにシード値を固定し，同じ画像が選ばれるようにした。

※ 5) pix2pix はコンテンツ生成型の AI 技術である GAN（敵対的生成ネットワーク）の一種である。画像 A と画像 B のペアで学習することで AtoB の画像変換が可能である。

※ 6) Grad-CAM では画像を入力した際の対象層の特徴量マップと勾配値を計算することで重みづけを行い，分類を行う際に，どのフィルタを重視したのかを計算する。フィルタ枚数分の結果を合計し正規化することで，AI 注目領域を可視化することができる。本研究では図２で示されるネットワーク中の矢印で示す ReLu 層（6*10 サイズ，フィルタ 128 枚）における出力を行う。なお事前検証において 4200 枚通常画像学習モデルにおいてスケッチ・注視点情報との比較結果を以下に示す。



Grad-CAM と注視点画像・スケッチ画像の比較

参考文献

01) 山田悟史，大野耕太郎：Deep Learning を用いた印象評価推定 AI の作成と検証，日本建築学会計画系論文集，第 84 巻，第 759 号，pp987-995, 2019.5

02) Hina Kinugawa, Atsushi Takizawa：Deep Learning Model for Predicting Preference of Space by Estimating the Depth Information of Space using Omnidirectional Images, Proceedings of the 37th eCAADe and 23rd SIGraDi Conference, 2, Porto, Portugal, pp.61-68, 2019.9

03) Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra:Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,arXiv:1610.02391, 2016.10

04) Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros:Image-to-Image Translation with Conditional Adversarial Networks,arXiv:1611.07004, 2016.11