視線解析および深層学習に基づく印象評価の可能性
－ 木造住宅密集地域における街路の魅力評価を例に －

沖　拓弥＊

Possibility of Evaluating Impressions Based on Gaze Analysis and Deep Learning
- A Case Study of Attractiveness Evaluation of Streets in Densely Built-up Wooden Residential Area -

Takuya Oki

This paper aims to investigate the possibility of impression evaluation based on gaze analysis of subjects and deep learning, using an example of evaluating street attractiveness in densely built-up wooden residential areas. Firstly, the relationship between the subjects' gazing tendency and their evaluation of street image attractiveness on the monitor is analyzed by measuring the subjects' gaze with an eye tracker. Next, we construct a model that can estimate an attractiveness evaluation result using convolutional neural networks (CNNs). Besides, using the method of gradient-weighted class activation mapping (Grad-CAM), we attempt to visualize which street components can contribute to the attractiveness evaluation. Finally, we discuss the similarity between the subjects' gaze tendencies and activation heatmaps created by Grad-CAM.

Keywords: 視線解析（gaze analysis），畳み込みニューラルネットワーク（convolutional neural network），Grad-CAM（Grad-CAM），木造住宅密集地域（densely built-up wooden residential area），魅力評価（attractiveness evaluation）

1. はじめに

本稿は，木造住宅密集地域における街路の魅力評価を例に，被験者の視線解析および深層学習に基づく印象評価の可能性について検討することを目的とする．具体的には，まず，モニタ上の街路画像の魅力を評価している被験者の視線をアイトラッカーで計測し，被験者の注視傾向と魅力評価の関係を分析する．次に，畳み込みニューラルネットワーク（CNN）（Krizhevsky, et. al., 2012）を用いた印象評価推定モデルを構築し，街路のどの構成要素が魅力評価に寄与しうるかを Grad-CAM（Selvaraju, et. al., 2016）により可視化するとともに，被験者の注視傾向との類似性について考察する．

2. 魅力評価時における視線解析の流れ

2.1. 視線解析の意義

人々が対象から受ける印象を調査・分析する上では，SD（Semantic Differential）法（Osgood, et. al, 1957）などの主観的評価手法が用いられる場合が多い．一方で近年，様々な計測技術が急速に進歩しており，こうした技術を活用した客観的な評価手法を構築することの意義は大きい．

2.2. 視線計測の方法

モニタ上に提示された画像を見る被験者の視線を計測することから，スクリーンベースの小型アイトラッカーである Tobii Pro ナノ（トビー・テクノロジー，2020）を用いる（図 1）．被験者の着座位置とモニタとの間の距離は約 60cm であり，視線のキャリブレーション後に計測を開始し，前稿（木澤・沖，2020）で述べた一連の印象評価アンケートに回答してもらう．視線計測ログとアンケート回答ログは，時刻情報に基づいて実験終了後に対応付ける．

2.3. 注視時点の特定方法

視線計測ログは約 1/60 秒ごとに記録されるが，ここには視線が大きく動いている状態（saccade）とほぼ停留している状態（fixation）の両方が含ま

＊　正会員　東京工業大学環境・社会理工学院（Tokyo Institute of Technology）
〒 152-8550　東京都目黒区大岡山 2-12-1　E-mail: oki.t.ab@m.titech.ac.jp

れる．しかし，被験者が魅力評価を行うためには，画像中の対象を注視し，それが何であるかを認識することが必要である．そこで以降の分析では，対象を注視していると見なせる時点の視線情報のみを用いる．

注視点には状況に応じた様々な定義が存在する（図 2(a)）．ただし，本実験条件下では眼球運動速度が全体的に遅いため（図 2(b)），試行的に「35pixel 以内の範囲に 100ms 以上視線が停留している場合」を注視時点と見なす．

2.4. 注視時間密度の算出方法

事前にキャリブレーションを行っても，視点座標にはある程度の誤差が含まれる．そこで，注視時間の密度分布を求める際には，前節で抽出した注視点群の座標値そのものではなく，注視点群にカーネル密度関数を適用し平滑化する方法を採用した（図 3）．

2.5.Semantic segmentation 結果と注視時間の対応

Google Street View API で取得した木密地域の街路画像（計 100 枚，384pixel × 513pixel）の各 pixel を，深層学習を用いた Semantic Segmentation により，19 分類 [1] のいずれかに分類した．これを図 4 に示す方法で 100pixel × 100pixel の注視時間密度分布と対応付けることで，各画像における分類別の注視時間を算出する．このとき，画像中で占める面積の大きい分類ほど，注視時間が長く算出される傾向が見られた．そこで，注視時間の分類別構成比は，各分類の注視時間をそれぞれの占有面積で基準化した上で算出する．

3. 魅力評価時における視線解析の結果

3.1. 被験者実験の概要

前稿（木澤・沖，2020）で述べた木密地域の印象評価アンケートの回答者 32 名のうち，30 名の視線を計測した（30 名×写真 10 枚＝ 300 組）．写真 1 枚につき 18 個の設問があり，うち 5 問が魅力評価に関する設問である（表 1 の⑭～⑱）．以下では，前稿（木澤・沖，2020）で分析対象とした 279 組に含まれ，かつ，視線ログが正常に記録できている計 211 組の視線ログデータを用いて，被験者の注視傾向などを分析する．

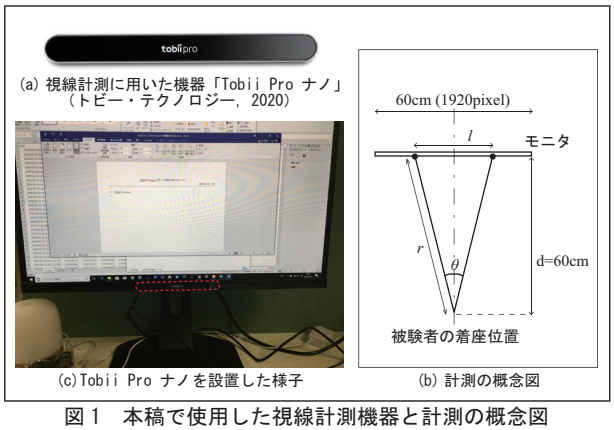


図 1　本稿で使用した視線計測機器と計測の概念図

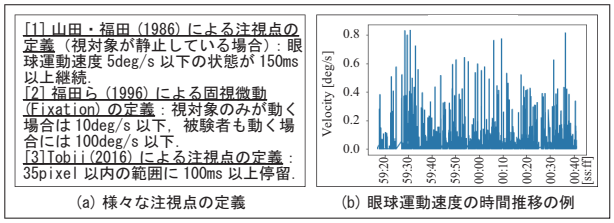


図 2　注視時点の特定方法と眼球運動速度

図 3　注視時間の密度分布

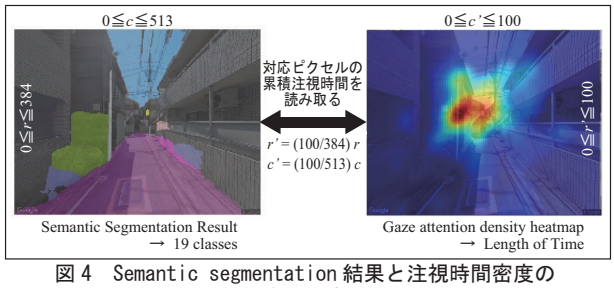


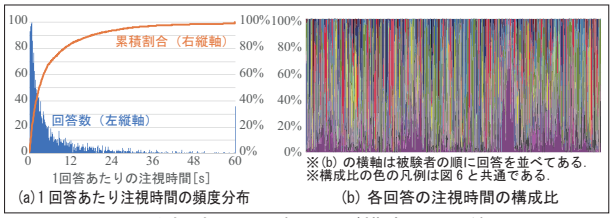
図 4　Semantic segmentation 結果と注視時間密度の対応付け方法

表 1　質問項目の一覧

【景観評価に関する設問】 ①開放的な/閉鎖的な，②親しみのある/疎外感のある，③活気のある/閑散とした，④快適な/不快な，⑤緑の豊かな/緑が乏しい，⑥落ち着いた/落ち着かない，⑦明るい/暗い，⑧昔ながらの/現代的な，⑨安心感のある/安心感のない，⑩すっきりした/ごみごみした，⑪生活感のある/生活感のない，⑫居心地のいい/居心地の悪い，⑬清潔な/汚い． 【魅力評価に関する設問】 ⑭好き/嫌い，⑮面白い/つまらない，⑯住みたい/住みたくない，⑰通りたい/通りたくない，⑱魅力的な/魅力的でない．

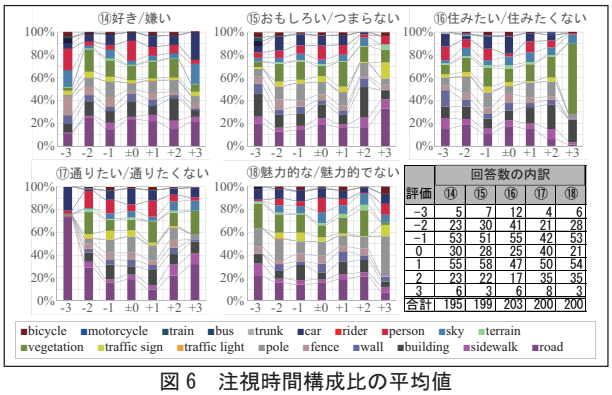
3.2. 注視時間とその内訳

設問 1 問の回答に要した注視時間の平均は 7.2 秒であり，回答全体の約 8 割の注視時間が 10 秒以下であった（図 5(a)）．2.4 節の方法で推定し



(a) 1 回答あたり注視時間の頻度分布　(b) 各回答の注視時間の構成比

図 5　注視時間の分布および構成比（回答別）



評価	⑭	⑮	⑯	⑰	⑱
−3	5	7	12	4	6
−2	23	30	41	21	28
−1	53	51	55	42	53
0	30	28	25	40	21
1	55	58	47	50	54
2	23	22	17	35	35
3	6	3	6	8	3
合計	195	199	203	200	200

図 6　注視時間構成比の平均値

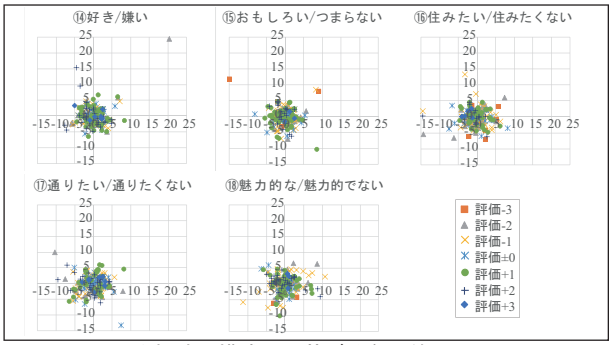


図 7　注視時間構成比に基づく各回答の位置関係

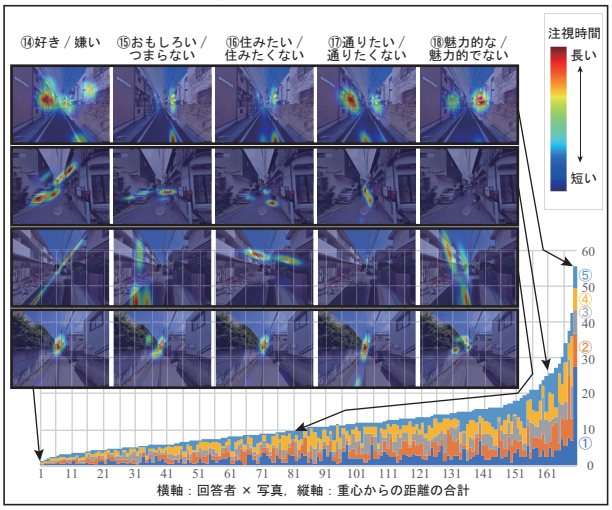


図 8　各回答の重心からの距離と注視時間密度分布との関係

た分類別の注視時間に着目すると（図 5(b)），回答ごとに注視する対象は大きく異なることがわかる．ただし平均値を見ると，設問や評価値に特有の注視傾向は見られない（図 6）．

3.3. 同一写真回答時における設問ごとの注視傾向

回答ごとに得られている 19 分類それぞれの注視時間構成比をもとに，多次元尺度構成法（MDS）を用いることで，各回答間の位置関係を 2 次元平面上にプロットできる（図 7）．

この XY 座標をもとに，各被験者による同一写真に対する設問⑭～⑱の重心座標を求め，各設問と重心座標との間の距離を積み上げ棒グラフで示した（図 8）．棒が長いほど，注視時間構成比の設問間の傾向の違いが大きいことを意味している。このことは，実際の注視時間密度分布を見ても確認できる．

4. 深層学習に基づく印象評価結果の分析

4.1. 印象評価推定モデルの概要

本章では，被験者アンケートや視線計測時と同じ街路画像を用いて，各画像に対する①～⑱の各印象評価値を推定するモデルを構築する．具体的には，街路画像と各被験者による印象評価結果（-3 ～ +3 の 7 クラス）のペアを学習データとして，教師あり学習の一種である CNN で評価項目ごとに学習させる．CNN は，畳込み層とプーリング層と呼ばれる 2 種類の層を交互に積み重ねた多層ニューラルネットワークである．本稿では，クラス数は多くないものの計算リソースに限りがあることから，文献（チームカルポ，2019；ショレ，2018）を参考に 6 層の畳込み層を用いる（表 2）．また，CNN の可視化手法の一つである

表 2　本稿で用いる畳込みニューラルネットワークの構造

Layer	Input shape	Output shape
Input（入力層）	120 x 120 x 3 (RGB)	120 x 120 x 32
Convolutional layer 1（畳込み層 1）	120 x 120 x 32	120 x 120 x 32
Batch normalization 1（正規化 1）	120 x 120 x 32	120 x 120 x 32
Convolutional layer 2（畳込み層 2）	120 x 120 x 32	120 x 120 x 32
Batch normalization 2（正規化 2）	120 x 120 x 32	120 x 120 x 32
Max pooling 1（プーリング層 1）	120 x 120 x 32	60 x 60 x 32
Dropout 1（ドロップアウト層 1）	60 x 60 x 32	60 x 60 x 32
Convolutional layer 3（畳込み層 3）	60 x 60 x 32	60 x 60 x 64
Batch normalization 3（正規化 3）	60 x 60 x 64	60 x 60 x 64
Convolutional layer 4（畳込み層 4）	60 x 60 x 64	60 x 60 x 64
Batch normalization 4（正規化 4）	60 x 60 x 64	60 x 60 x 64
Max pooling 2（プーリング層 2）	60 x 60 x 64	30 x 30 x 64
Dropout 2（ドロップアウト層 2）	30 x 30 x 64	30 x 30 x 64
Convolutional layer 5（畳込み層 5）	30 x 30 x 64	30 x 30 x 128
Batch normalization 5（正規化 5）	30 x 30 x 128	30 x 30 x 128
Convolutional layer 6（畳込み層 6）	30 x 30 x 128	30 x 30 x 128
Batch normalization 6（正規化 6）	30 x 30 x 128	30 x 30 x 128
Max pooling 3（プーリング層 3）	30 x 30 x 128	15 x 15 x 128
Dropout 3（ドロップアウト層 3）	15 x 15 x 128	15 x 15 x 128
Flatten（平滑化層）	15 x 15 x 128	28800
Output（出力層）	28800	Num. of classes

表 3　印象評価モデルの推定精度（評価項目別．−3, −2, −1, 0, +1, +2, +3 の 7 クラス分類）

	開	親	活	快	緑	落	明	昔	安	す	生	居	清	好	面	住	通	魅
	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	⑪	⑫	⑬	⑭	⑮	⑯	⑰	⑱
適合率（平均）	28.9%	28.7%	42.2%	30.0%	43.0%	46.9%	27.4%	42.5%	16.8%	31.1%	18.2%	45.0%	56.3%	25.6%	22.6%	27.2%	31.9%	24.5%
再現率（平均）	25.1%	20.2%	22.3%	19.3%	24.8%	16.7%	21.0%	15.5%	17.5%	23.9%	16.8%	25.5%	23.7%	24.5%	14.1%	25.5%	22.5%	31.3%
F値	26.9%	23.7%	29.2%	23.5%	31.5%	24.6%	23.8%	22.7%	17.1%	27.0%	17.5%	32.6%	33.4%	25.1%	17.4%	26.3%	26.4%	27.5%
総合精度	37.5%	36.9%	39.4%	38.5%	32.3%	46.9%	36.9%	35.4%	43.1%	41.5%	40.0%	43.8%	44.6%	38.5%	29.2%	35.4%	30.8%	33.8%
最多クラス割合	28.1%	38.5%	36.4%	35.4%	27.7%	46.9%	30.8%	35.4%	36.9%	32.3%	38.5%	32.8%	32.3%	27.7%	30.8%	30.8%	30.8%	30.8%
最多クラスNo.	+1	+1	-1	+1	+1	+1	+1	+1	+1	+1	+1	+1	-1	+1	-1	-1	+1	-1

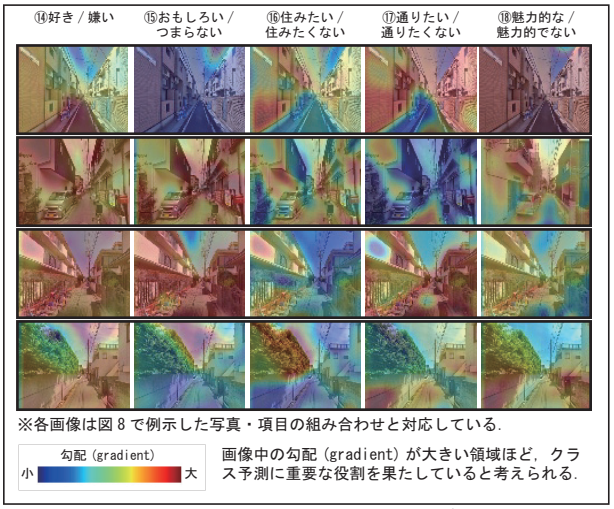


図 9　Grad-CAM による活性化ヒートマップの可視化例

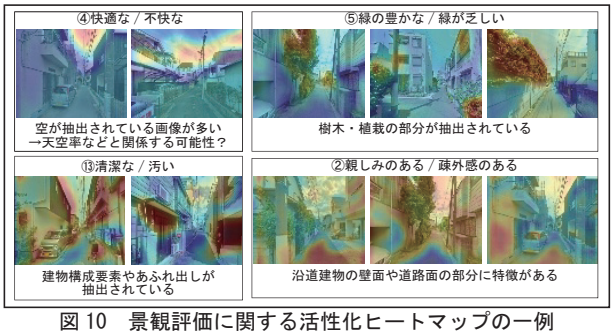


図 10　景観評価に関する活性化ヒートマップの一例

Grad-CAM で活性化ヒートマップを作成し，これをもとに各モデルの判断構造の理解を試みる．

4.2. 印象評価推定モデルの学習結果

学習データを教師データ 6 割，検証データ 2 割，テストデータ 2 割となるように分けた上で，印象評価推定モデルに学習させ，学習済みモデルをテストデータに適用した場合の推定精度を表 3 に示してある．適合率と再現率の調和平均である F 値は概ね 3 割前後，総合精度（accuracy）は 4 割前後であり，モデルの精度はまだ十分とは言い難い．これは，学習データ数が少ないことや，正解データが特定の評価値（+1 や -1）の回答に偏っていること，さらには同一の画像・設問に対しても評価の個人差が大きいことなどが原因と考えられる（本モデルでは現状，評価の個人差は考慮できない）．

一方で，Grad-CAM により可視化した活性化ヒートマップを見ると（図 9），CNN によって画像の特徴が捉えられている可能性のある，興味深い結果も見受けられる．例えば，⑯は街路の沿道建物の構成要素が重視される傾向が見られるのに対し，⑰は道路部分とそれ以外の部分の勾配の差が比較的明確に表れている．また，景観評価に関する活性化ヒートマップを見ても，特徴抽出に成功しているケースがある．

ここで，注視時間密度分布（図 8）と活性化ヒートマップ（図 9）とを比較したが，両者の間に類似性はあまり見られないことがわかる．より詳細に，テストデータとして用いた全画像について，ピクセル単位で両者を比較してみたが，コサイン類似度は最大でも 0.7 程度，平均で 0.24 であった．この結果は，印象評価の際の注視点や評価構造が，人間（被験者）と人工知能（AI）で異なっている可能性を定量的に裏付けている．

5. まとめ

本稿では，木造住宅密集地域における街路の魅力評価を例に，まず，被験者の視線計測から得られる注視時間の分類別構成比や密度分布と，評価項目やその評価値との間の関係を分析した．次に，街路画像から被験者の魅力評価を推定する深層学習モデルを構築し，その推定精度を評価するとともに，Grad-CAM によりモデルの判断構造を可視化し，その特徴について考察した．

本稿の一連の分析の結果からは，被験者の注視傾向と印象評価の間には明確な関係性は見い

だせず，また，精度の高い印象評価推定モデルを構築することはできていない．この原因の一つとして，印象評価における個人差の影響が挙げられる．個人差の影響は，被験者アンケートから直接得られる評価値だけでなく，被験者ごと・写真ごと・設問ごとの注視傾向の違いにも顕著に表れている．一方で，本稿で構築した深層学習ベースの印象評価推定モデルでは，個人差を含む学習データを用いたものの，構造上，個人差を考慮可能なモデルとはなっていない．今後は，より多くの被験者データを収集しつつ，個人差の考慮方法について検討することで，視線計測および深層学習に基づく印象評価の可能性を引き続き追求する．

謝辞

本稿での視線計測は，木澤佐椰茄氏（東京工業大学環境・社会理工学院）の卒業論文研究におけるアンケート調査とあわせて行ったものである．計測にご協力いただいた方々に謝意を表します．

注

1）DeepLabV3+ モデル（Chen et. al., 2018）の学習に用いた CityScapes データセット（Cordts et. al., 2016）で定義されている以下の 19 分類に基づく：road, sidewalk, building, wall, fence, pole, traffic light, traffic sign, vegetation, terrain, sky, person, rider, car, trunk, bus, train, motorcycle, bicycle.

2）表 3 の適合率算出時には予測合計が 0 のクラスを，再現率算出時には正解合計が 0 のクラスを除いて平均値を求めている．

参考文献

木澤佐椰茄・沖拓弥（2020）木造住宅密集地域の魅力構造に関する基礎的分析 その 1 ：木造住宅密集地域における街路の特徴と印象評価の関係，「日本建築学会大会学術講演梗概集」，E-1，917-918.

ショレフランソワ（2018）「Python と Keras によるディープラーニング」，株式会社クイープ訳，巣籠悠輔監訳，マイナビ出版.

チームカルポ（2019）「必要な数学だけでわかるディープラーニングの理論と実装」，秀和システム.

トビー・テクノロジー（2020）Tobii Pro ナノ製品紹介ページ．（2020.3.24 参照）<https://www.tobiipro.com/ja/product-listing/tobii-pro-nano/>

福田亮子・佐久間美能留・中村悦夫・福田忠彦（1996）注視点の定義に関する実験的検討，「人間工学」，**32**(4)，197-204.

山田光穂・福田忠彦（1986）画像における注視点の定義と画像分析への応用，「電気通信学会論文誌 D」，**69**(9)，1335-1342.

L. Chen, et. al. (2018) Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, arXiv:1802.02611.

M. Cordts, et. al. (2016) The Cityscapes Dataset for Semantic Urban Scene Understanding. （2020.8.31 参照）<https://www.cityscapes-dataset.com>

A. Krizhevsky, I. Sutskever, and G. E. Hinto, (2012) ImageNet classifi cation with deep convolutional neural networks, *Advances in neural information processing systems*, **25**(2), 1097–1105.

C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, (1957) *Measurement of meaning*, Urbana, University of Illinois Press.

R. R. Selvaraju, et al. (2016) Grad-CAM: Visual explanations from deep networks via gradient-based localization, arXiv:1610.02391.

Tobii (2016) Tobii Studio User's Manual Version 3.4.5. （2020.3.25 参照） <https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-studio-user-manual.pdf>