風景写真における人間と AI の美的価値判断の比較

松田はるか（指導教員：佐藤有理）

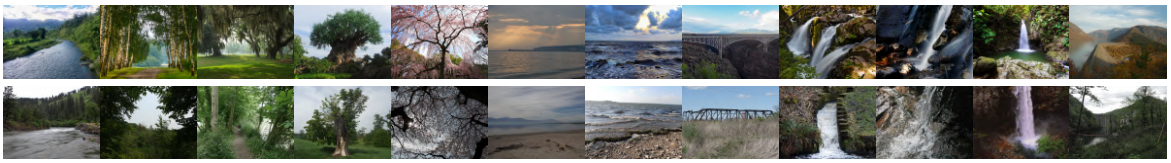

図 **1.** beauty ラベル写真（上段）と not-beauty ラベル写真（下段）

1 はじめに

デジタル環境の普及により、大量の写真データの整理が日常的な課題となっている。一部を保存し、一部を削除する際に人々が従う基準は何か。本研究は、風景写真の美しさを決定する基準に焦点を当てた。美学における「ピクチャレスク」は、構図などの非美的特性が美的価値を引き起こすことを示唆している（Gilpin, 1782; Sibley, 1959）。本研究はこの概念の問題に、人間と AI を比較する形でアプローチした。美しいと評価された風景写真を収集し（Section 2）、画像分類タスクを実施した。その結果、人間は美的特性に基づいて正確に画像を識別できた (Section 3.1) が、CNN モデルはチャンスレベルを超える結果を示さなかった (Section 3.2)。しかし、アイトラッキングと Grad-CAM によるアテンションマップ分析では、課題遂行時の注意の向け方において、人間とモデルに大きな差はなかった（Section 4）。一方で、few-shot learning を行った大規模言語モデル ChatGPT-4o は、人間と同等の高い分類精度を示した（Section 5）。これにより、美的価値判断における「知識」の重要性が示され、従来のピクチャレスク的見解に再考を促す結果となった。

2 データ収集

以降の実験で使用する写真データは、Flickr (https://www.flickr.com) から非営利目的で利用可能な画像を収集した。以下の手順で選別・調整を行った。

Step 1. 画像内容コントロール. Flickr で風景関連のキーワード（mountain, lake, field, bridge, tree）を検索し、内容が似た画像を 100 ペア選んだ。その後、ChatGPT を使って各画像の内容をテキスト化し、Semantic Textual Similarity (STS) を用いて画像間の内容的類似度を評価した。ペア間の意味的類似度が 0.6 以上となる画像ペアが 100 枚ずつできるまで、上の手続きを繰り返した。

Step 2. 人間によるアノテーション. Step 1 で得られた 100 ペアの画像について、オリジナル選定者一名以外に、4 名のアノテーターに各画像を評価してもらった。「美しい風景作品としてよくとれているものを選んでください」と教示した。一人 50 画像程度が与えられ、一画像につき 2 名分のアノテーションデータを得た。2 名のうち 1 名がオリジナル選定者と同評価だった画像を採用した。結果、この基準をクリアした写真ペアは 74 ペアだった。同評価にならない写真ペアの分（残りの 26 ペア分）については、step 1 から戻って画像データを入れ替え、アノテーターによる評価が基準をクリアできるまで繰り返した。

Step 3. 画像彩度コントロール. 構図の影響に焦点を当てるため、色、特に彩度をコントロール。Python と OpenCV を使い、美しいラベル画像に合わせて、対応する美しくないラベル画像の彩度を調整し、ペア間の類似度が 0.3 以上になるようにした。

3 画像分類課題

3.1 人間を対象にした実験

CrowdWorks および qualtrics のプラットフォームを用いて日本語話者 146 名を対象にオンラインで実施した。1 人 40 枚の画像を評価し、回答は一画像あたり平均 28.81 回分得られた。beauty ラベル写真での正答率は 80.2%、not-beuty ラベル写真での正答率は 47.2% だった。beauty ラベル写真のうち、正答率がチャンスレベル 50% と有意な差を持つものは 69 枚だった（フィッシャーの直接確率検定で 5% 基準）。他方で、not-beauty ラベル写真のうち、正答率がチャンスレベルと有意な差を持つものは 22 枚だった。ペアとして上記の基準を満たした写真は 9 ペアだった。直接確率検定の閾値を 10% に下げた場合には、基準を満たしたのは 12 ペアだった。図 1 にはその 12 ペア (合計 24) それぞれの写真画像を示す。以降の分析はこれらの写真を分析対象とした。これらの結果は、美しい風景作品としてよくとれている/とれていないと、人々が共通して判断する写真画像があることを示す。

3.2 機械学習モデル実験

OpenCV の標準的な技法でデータ拡張を行ったうえで、カテゴリごとに画像 12 枚 (×84) を、training9 枚、validation2 枚、test1 枚にデータ分割をした。cross-validation の方法を使い、さらに画像順をラン

ダムに変え、合計 48 の学習モデルを、VGG16 の pretraining 付きの CNN として Python Keras ライブラリで構築した。全体の正答率は、beauty ラベル写真で 50.0%、not-beauty ラベル写真で 37.5% だった。これらの結果は、美的価値判断について、画像情報だけで学習した CNN モデルは 50% チャンスレベル以上には判別できていないことを示す。

4 人間とモデルのアテンション比較

4.1 人間のアイトラッキング実験

日本語話者 16 名が参加し、Tobii Pro Spark を使用して視線データを収集した。通常のキャリブレーションのあとに、(1) 注視点 (*) が呈示されて、一定時間視線が集まったら、(2) 画像が 10 秒間呈示され、(3) 黒画面が呈示され、そのときに「美しい風景作品としてよくとれているか」を口頭で答えた。これを 12 トライアル繰り返した。正解した項目に限定し、画像ごとに二者間の画像類似度を OpenCV、scikit-image 上で計算した。色というよりは構造に注目しているので、structural similarity (SSIM) (Wang et al., 2004) を使った。Müller et al (2024) に従い、tobii のヒートマップ出力のオリジナルで得られるグレースケール画像を、バイナリ (50% を基準) に変換し、以降の類似度分析に使用した（図 2）。全体平均は、beauty ラベル画像は 0.807、not-beauty ラベル画像は 0.806 だった。

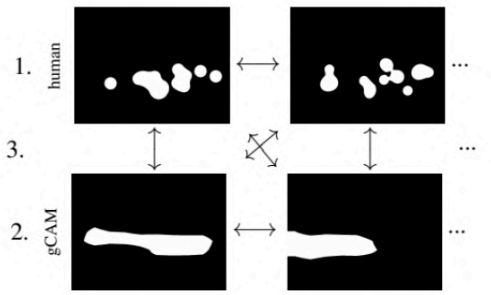

図 **2.** 人間と **AI** モデルの視線および注視領域の比較

4.2 GradCAM

画像分類課題のときに作った学習モデルと重みデータを使い、テスト画像の分類時のアテンションマップを GradCAM のフレームワークで作成した。Section 4.1 と同様に、バイナリに変換したものに類似性分析を行った。全体平均は、beauty ラベル画像は 0.847、not-beauty ラベル画像は 0.890 だった。

4.3 人間／モデル比較

人間と GradCAM それぞれのアテンションマップを写真ごとに二者間の画像類似度を計算した。全体平均は、beauty ラベル画像は 0.804, not-beauty ラベル画像は 0.798 だった。3 つの比較タイプの類似度データについて、beauty カテゴリーでは、統計的な有意差はなかった（統計量 $H = 4.25$, クラスカル・ウォリス検定)。not-beauty カテゴリーについては、正解を出力したモデルの数が極端に少なかった（サンプル数 5）ため統計検定は行うことは妥当でないと判断した。これらの結果は、人間とモデルの間の違いは、人間各々、モデル各々の違いと同程度であったということを示す。そのため、美的判断の画像分類課題において、人間とモデルのアテンションの向け方にそれほど大きな違いはないということが示唆される。

5 LLM ベース few-shot learning 分析

ChatGPT4o において、few-shot learning をしたうえで、Section 3.2 と同様の画像分類課題を行った。全体の正答率の結果（表 1 ）は、beauty ラベル画像は 91.7% で、not-beauty ラベル画像は 75.0% だった。画像上の情報だけで学習した CNN モデルと比べると、LLM ベース few-shot learning のパフォーマンスは高く、人間のパフォーマンスに匹敵するものだった。

表 1: 画像分類課題における正答率

モデル	美しい写真の正答率	美しくない写真の正答率
人間	82.9%	79.7%
CNN モデル	50.0%	37.5%
ChatGPT-4o	91.7%	75.0%

6 議論

我々に残された課題は、Section 3.2 での画像情報だけで学習した CNN モデルと Section 5 の few-shot learning 付き GPT4o モデルのギャップを埋めることである。具体的には、後者に加わる「知識」の内実の解明である。我々が Section 5 で行った分析は AI 心理学 (e.g., Binz & Schulz, 2023) に関連するもので、人間のタスクを LLM に実行させ、性能評価する形式である。しかし、モデルの学習する内容がブラックボックスであるため、認知科学としては十分とはいえない。今後は、ファインチューニングを通じて知識を段階的にモデルに与えるなどして、出力がどのように変化するかを分析することで、美的価値判断に寄与する要因をさらに明らかにしていく必要がある。

参考文献

[1] Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120.

[2] Gilpin, W. (1782/1800). *Observations on the River Wye and Several Parts of South Wales, &c Relative Chiefly to Picturesque Beauty; Made in the Summer of the Year of 1770* (5th ed.). London: Cadell & Davies.

[3] Müller, R., Duerschmidt, M., Ullrich, J., Knoll, C., Weber, S., & Seitz, S. (2024). Do humans and convolutional neural networks attend to similar areas during scene classification: Effects of task and image type. *Applied Sciences*, 14(6), 2648.

[4] Sibley, F. (1959). Aesthetic concepts. *The Philosophical Review*, 68, 421–450.

[5] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13, 600–612.