D-017

クラウドソーシングを用いた画像の美的評価に関する検討

A Note on Aesthetic Evaluation of Images Using Crowdsourcing

嘉藤 悠大† 桂井 麻里衣† 田島敬史‡
Yudai Kato Marie Katsurai Keishi Tajima

1. はじめに

SNS 上の投稿画像に対する感情分析や美的評価は，データマイニングや情報推薦を高度化する技術として活発に研究開発されている．従来研究では，明度や構図などを表す画像特徴が人手により設計されてきた．最近の研究では，画像と主観評価ラベルを用いた教師あり学習により，感情や美しさに関連する特徴を自動で抽出する手法が主流となりつつある．しかし，これらの機械学習アプローチは，正確な教師ラベルをもつ大量のデータが必要という問題点がある．信頼性の低いデータや偏ったデータで学習した場合，テストデータの分析性能に影響を及ぼす．また，データ数の不足から過学習などの問題が起きることもある．

そこで，様々な分野においてデータの収集に利用されているクラウドソーシングを用い，画像に対する主観評価を大量に収集することを考える．クラウドソーシングは，インターネット上などで不特定多数の人々を募って，なんらかのタスクを依頼するタスク実施形態であり，安価かつ短時間で大量のタスクを処理可能という利点がある．しかし，中には大量のタスクを処理するために，指示を読まずにタスクを進めてしまうスパマと呼ばれるワーカが存在する．そのようなワーカによる不適切なタスク結果を排除するために，同じタスクを複数のワーカに重複して割り当て，それらの結果の多数決などを用いて不適切な結果を識別する手法がよく用いられる．しかし，それらの手法は，適切な結果というものが客観的に決まるようなタスクにのみ有効であり，主観評価という個人差の大きなタスクにおいては有効ではない．そのため，そのようなタスクにおいて，質の良いデータを収集する新たな方法論が求められる．

本研究では，主観評価のクラウドソーシングにおいてスパマの影響を低減させ，高品質なデータを収集する方法を検討する．特に評価タスク時にワーカに与える情報と受注資格の設定方法に着目し，美的評価モデルの学習用データセットとの相関係数を算出することでクラウドソーシング成果物の品質を評価する．まず実験 1 では受注資格の設定方法が評価の品質に与える影響を調査した．実験 2 ではワーカに与える情報が美的評価に与える影響を調査した．

2. プラットフォームとデータセット

本章では画像の美的評価をクラウドソーシングする際の受注資格の好ましい設定条件を実験結果にもとづいて提案する．プラットフォームとして，現在デファクトスタンダードとなっている Amazon Mechanical Turk[1] (MTurk)を用いる．MTurk では，一つの仕事を Human Intelligence Task (HIT)と呼ぶ．HIT は短時間で終わる簡単なタスクであることが多い．依頼者には HIT の品質を向上させるため，HIT の受注資格を設定する機能や，提出された HIT を拒否する権利が与えられている．受注資格は MTurk が予め用意したものを選択するほか，依頼者自身が作成することもできる．受注資格によっては追加料金が必要となる．

提案タスクの画像データは AVA データセット [1] から取得した．AVA データセットは DPchallenge[2]から得られた大規模な美的評価に関するデータセットである．DPchallenge とは，プロ，アマ問わず写真家が Challenge と呼ばれるお題付きのコンテストに写真を提出し，お互いの写真に 1 から 10 のスコアをつけて，写真の美しさを競い合うウェブサイトである．コンテストの参加者が評価者も兼ねるため，写真に造詣が深いユーザが多いという点から，他のソーシャルメディアのデータに比べて評価の信頼性が高いといえる．AVA データセットには 255,530 枚の画像が含まれており，各画像には 78 件から 549 件，平均 210 件の評価が付与されている．従来研究では各画像に与えられた評価の平均を単一の美的スコアへと集約することが多いため，本研究もそれにならい，評価の平均をスコアに用いる．

3. 実験 1

HIT の受注資格が美的評価収集に与える影響の調査を目的とし，画像一枚につき 50 名のワーカから評価を収集した．

3.1 画像の抽出

AVA データセットに含まれている画像全てについて評価を収集することはコスト面から難しい．そのため，AVA データセットのスコア分布に基づきランダムに 20 枚の画像を抽出した．具体的には，最小値 1，最大値 10，階級幅 1 のヒストグラムを生成し，これを正規化してスコアの確率分布とみなすことで，各階級からの抽出枚数を重み付けした．

3.2 評価画面の設計

DPchallenge は評価者に対し画像が Challenge（お題）に合っているかどうかを考慮するよう指示しており，かつ各画像には撮影者がタイトルを付けている．このことから，AVA データセットのスコアは Challenge やタイトルが評価に影響している可能性があると考えた．より AVA データセットに近い状況で美的評価を収集するため，評価画面に Challenge，タイトル，画像を提示し，画像が Challenge に合っているかどうかを考慮する旨の指示文を挿入した．

3.3 受注資格

本実験で用いた受注資格を表 1 に示す．承認率は提出した HIT が承認された割合を表し，承認率の低いワーカはスパマである可能性が高いといえる．また，承認数は承認さ

† 同志社大学 Doshisha University
‡ 京都大学 Kyoto University

[1] https://www.mturk.com

[2] https://www.dpchallenge.com

表 1　受注資格の条件一覧.

条件	承認数	承認率
1	100 以上	95 以上
2	なし	95 未満
3	100 未満	なし
4	5,000 以上	98 以上
5	なし	なし

れた HIT 数を表す．承認数が多いほど MTurk に慣れているワーカであり，少ないほど新しいワーカであるといえる．条件 1 はクラウドソーシングの研究でよく用いられる設定である [2]．条件 2 はスパマである可能性が高いワーカを，条件 3 では新しいワーカを多く含むと考えられる．条件 4 は MTurk に慣れた質の良いワーカを多く集めることを目的に，承認数と承認率を大きく制限した．条件 5 は何も受注資格を設定しなかった場合を指す．

各条件における結果の再現性を確認するため，同条件での実験を二回ずつ実施した．いずれも期限は一週間に設定した．なお，HIT の言語を英語とする場合，US 在住のワーカに限定しなければ成果物の品質が低下するという報告 [3] があるため，すべての条件において US 在住のワーカのみに受注資格を与えた． また，全ての実験において同じ ID を持つワーカが複数回実験に参加することのできないように設定した．

3.4 実験結果

各条件で得られた評価を画像ごとに集約し，クラウドソーシングによる美的スコアを算出した．得られたスコアの評価として，AVA データセットにおけるスコアとの相関係数および Mean Absolute Error (MAE) を算出した結果を表 2 に示す．なお条件 2 の 2 回目の実験のみ期限内に回答が集まらなかったため，45 名分のみを分析した．1 回目の実験では条件 4 のみ AVA データセットのスコアと高い相関を示した (p<0.05)．2 回目の実験では条件 2, 3, 4 で相関があった (p<0.05)．最も質の良いワーカが得られることを期待していた条件 4 では，やはり最も AVA データセットに近い結果が得られた．

次に，各受注条件における二回の実験結果の比較を表 3 に示す．二回分の結果が近いほど，クラウドソーシングによるデータ収集の再現性が期待できる．全ての条件において成果物の間に相関があった (p<0.05)．特に条件 2, 3, 4 においては 0.8 以上の相関係数を示したため，これらの条件は再現性が高いと考えられる．

表 2，表 3 の結果から各条件を総合的に評価すると，条件 4 で最も良い結果が得られることがわかった．そして，条件 1 は条件 2, 3 よりも悪い結果を示した．このことから，承認率，承認数の制約を厳しくすることは必ずしも良い結果を導かないことがわかった．特に，条件 1 の結果が条件 5（受注資格を設定しない場合）の結果と大きく変わらなかったことからも，条件 1 は主観評価に不十分であることがわかる．

表 2　各受注条件における AVA データセットとの比較.

条件	1 回目		2 回目	
	相関係数	MAE	相関係数	MAE
1	0.17	1.19	0.37	1.08
2	0.42	1.01	0.45	1.13
3	0.39	1.11	**0.54**	0.88
4	**0.55**	**0.94**	0.52	**0.77**
5	0.25	1.20	0.29	1.19

表 3　各受注条件における二回の実験結果の比較.

条件	相関係数	MAE
1	0.57	0.75
2	**0.93**	0.41
3	0.88	0.52
4	0.84	0.57
5	0.62	**0.33**

表 4 各受注条件における評価収集にかかった時間

条件	1 回目	2 回目
1	38 分	54 分
2	1 日 12 時間 34 分	期限切れ
3	5 時間 47 分	10 時間 44 分
4	30 分	41 分
5	44 分	53 分

各受注条件における評価収集に要した時間を表 4 に示す．評価の収集にかかる時間が長いほど，受注資格を満たすアクティブなワーカは少ないと考えられる．しかし，条件 4 は最も厳しい条件であるものの比較的短い時間で完了し，それと比較して条件 2, 3 は極めて長い時間を要した．今後の研究では，複数条件で発注日時を揃え，さらなる分析を行う予定である．

4. 実験 2

ワーカに与える情報と HIT の成果物の関係を調査するための実験を行った．具体的には，AVA データセットのスコアに Challenge とタイトルが影響を及ぼしているか否かを検証するために，Challenge とタイトルの提示の有無による評価収集結果を比較した．

表 5　提示する情報を変化させたときの AVA データセットとの比較.

条件	相関係数	MAE
A	**0.50**	1.23
B	0.47	**0.96**

表 6　条件 A と条件 B の比較.

相関係数	MAE
0.94	0.48

Challenge: Shoes
Title: Hail to the Kings, baby
AVA データセット: 3.7
画像のみ: 2.6
タイトルあり: 3.8

Challenge: Wheres_Waldo_IV
Title: No Stopping Waldo
AVA データセット: 5.6
画像のみ: 6.9
タイトルあり: 5.5

Challenge: Late_Night
Title: Early sunrise during spring time
AVA データセット: 5.7
画像のみ:7.5
タイトルあり: 6.6

Challenge: Fantasy_World
Title: Imagining the Summit
AVA データセット: 4.1
画像のみ: 3.2
タイトルあり: 4.0

図 1　条件 A，B で平均に有意差のあった画像.

4.1 実験条件

本実験では，ワーカに画像のみを提示する場合を条件 A，画像とともに Challenge とタイトルを提示する場合を条件 B とよぶ．なお，条件 A, B ともに表 1 の条件 4 を適用した．

4.2 実験結果

各条件での結果と AVA データセットとの比較を表 5 に示す．相関係数は条件 A の方がやや高く，MAE は条件 B の方が良いということがわかる．また，条件 A, B による成果物の比較結果を表 6 に示す．相関係数が非常に高く，MAE も他実験に比べ低い値を示したことから，Challenge とタイトルの提示の有無は主観評価に大きく影響を与えないと類推できる．そこで，条件 A, B の評価間で Welch の t 検定を行ったところ，20 枚中 4 枚の画像で平均に有意差が認められた (p<0.05)．有意差が認められた画像を図 1 に示す．図において，AVA データセットに収録されていた評価スコア，画像のみを提示した実験で得られた評価スコア，タイトルも提示した実験による評価スコアをそれぞれ示す．これら 4 枚の画像に関しては，タイトルと Challenge を提示した条件 B の方が AVA データセットに近い値が得られた．したがって，画像とタイトルの組合せが主観評価に影響を及ぼす場合があることも考慮する必要があるといえる．

5. まとめ

本研究では， MTurk を用いた画像の主観評価収集において，ワーカの受注条件が及ぼす影響を調査した．20 枚の画像に対する実験結果から，高い承認数と承認率をワーカのみに受注資格を与えることの有効性が示された．特に，従来研究で用いられている承認率 95%以上，承認数 100 以上の設定では，主観評価の HIT には不十分であることが示唆された．また，画像のみならずタイトルと Challenge をともに示すことで，一部の画像の主観評価に影響を及ぼすことが示された．このことから AVA データセットのスコアは純粋な美的評価を示していない可能性があると考えられる．

本実験では画像に対する絶対評価を依頼したが，今後は複数画像間での相対評価を用いた美的評価の収集や，スパマの影響を低減させるような資格試験を提案する予定である．また，主観評価以外のタスクでも同様の実験を行うことで，主観評価特有の問題点を提起する．加えて，クラウドソーシングにかかる時間と受注条件の関係に関する詳細な分析を行う予定である．

参考文献

[1] Murray, N., Marchesotti, L. and Perronnin, F., “AVA: A large-scale data base for aesthetic visual analysis”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2408–2415, 2012.
[2] Robinson, J., Rosenzweig, C., Moss, A. J. and Litman, L., “Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool”, PLOS ONE, Vol.14, No. 12, 2019.
[3] Goodman, J. K., Cryder, C. E. and Cheema, A., “Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples”, Journal of Behavioral Decision Making, Vol. 26, No. 3, pp. 213–224, 2013.