画像特徴のクラスタリングによる形容詞の視覚的な多様性の分析

田中 優衣[1,a)]　カストナー マークアウレル[2,b)]　川西 康友[3,1,c)]　駒水 孝裕[1,d)]　井手 一郎[1,e)]

概要

本研究では，形容詞を対象として，関連する画像の色やテクスチャなどの画像特徴の個々の多様性と人間が感じる視覚的多様性との関係を分析し，視覚的多様性の推定に有用な画像特徴を明らかにすることを目的とする．具体的には，アンケートによって人間が感じる視覚的多様性を調査するとともに，画像データセットを用いて画像特徴の多様性を定量化する．また，それらの結果に基づいて，画像特徴の多様性から視覚的多様性の大小を推定し，有効な画像特徴の組合せを特徴選択によって調べることで，各画像特徴と視覚的な多様性の関係を分析する．

1. はじめに

単語の視覚的多様性 [5] とは，ある単語から想起される心的イメージのばらつき度合いであり，言語とその視覚的なイメージの関係を探る上で重要な手がかりの 1 つである．例として，「乗り物」と「自動車」という 2 つの名詞では，前者に対応するイメージには様々な形が対応するが，後者に対応するイメージはより限られた形であるため，「乗り物」は「自動車」よりも視覚的多様性が高いと考えられる．また，図 1 に，「古い」と「黄色い」という 2 つの形容詞の例を挙げる．このとき，前者に対応するイメージには様々な色や形が対応する．一方，後者にも様々な形が対応するが，色は黄色に限られるため，人が感じるばらつき度合いは前者より小さいことが予想される．そのため，「古い」は「黄色い」よりも視覚的多様性が高いと考えられる．

単語の視覚的多様性に関する大規模なデータセットは存在せず，多数の単語について視覚的多様性を人手で定量化するためには多くのコストを要するため，機械的に推定することが有効であると考えられる．その推定手法として

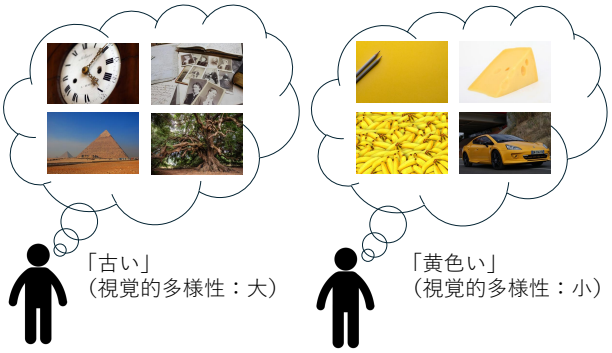

図 1: 単語の視覚的多様性の違いを表す例.

は，単語に関連する多数の画像の画像特徴に基づくことが考えられる．しかし，視覚的多様性は色や形など多くの要素と関連すると考えられるため，推定に有効な画像特徴は明らかではない．

そこで，本研究では，画像特徴の多様性に基づいて，形容詞に対して人間が感じる視覚的多様性の大小を推定する実験を行ない，個々の画像特徴と視覚的な多様性の関係を分析する．この分析により，人間の感覚に即した方法で単語の視覚的多様性を定量的に推定できるようにするほか，人間が感じる視覚的多様性に関する理解を深める一助となることが期待される．

本研究に関連する研究として，Yanai ら [12] は，Web 上の画像から形容詞の視覚性（Visualness）を算出する方法を提案した．視覚性は，ある概念がもつ視覚的な特徴の程度を表す尺度であり，画像アノテーション時の単語選択補助などを目的として提案された．また，Kastner ら [5] は，単語の視覚的多様性について調査した．類義語である 25 語の名詞について，Web 検索におけるヒット数に基づいて再構築した画像コーパス内の画像から特徴ベクトルを作成し，クラスタリングによって生成されたクラスタ数で名詞を順位付けした．しかし，これらは複数の画像特徴を統合して分析しているため，個々の画像特徴の多様性と人間が感じる視覚的多様性との関係は明らかではない．

そこで本研究では，画像特徴の多様性から視覚的多様性の大小を推定する際の特徴選択を通して，個々の画像特徴

[1] 名古屋大学
[2] 広島市立大学
[3] 理化学研究所
[a)] tanakay@cs.is.i.nagoya-u.ac.jp
[b)] mkastner@hiroshima-cu.ac.jp
[c)] yasutomo.kawanishi@riken.jp
[d)] taka-coma@acm.org
[e)] ide@i.nagoya-u.ac.jp

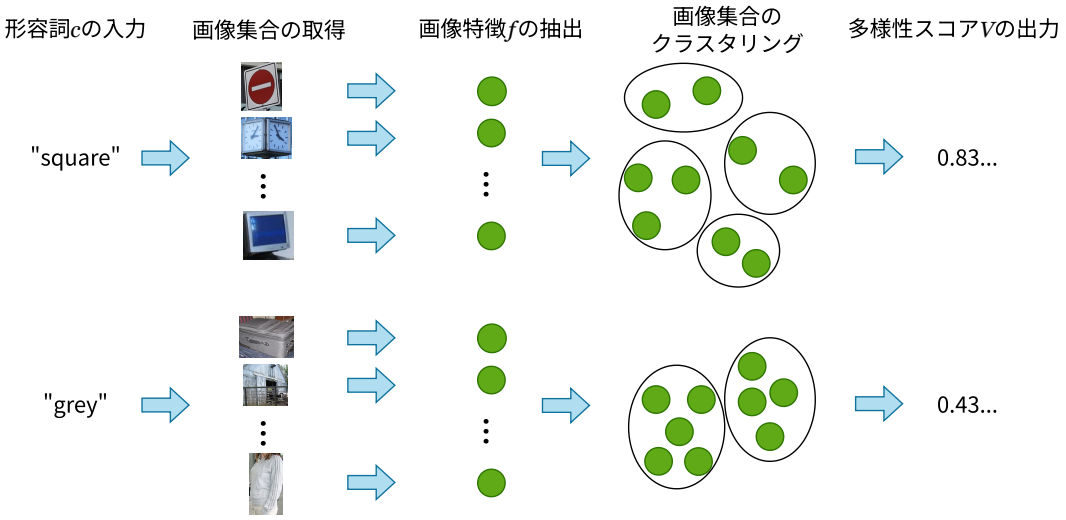

図 2: 画像特徴の多様性スコアを算出する手法の概要

と視覚的な多様性の関係を分析する．実験では，アンケートによって人間が感じる視覚的多様性を調査するとともに，画像データセットを用いて画像特徴の多様性を定量化する．また，それらの結果に基づいて，画像特徴の多様性から視覚的多様性の大小を推定し，各画像特徴と視覚的な多様性の関係を分析する．その結果，画像特徴の多様性を用いて，高い精度で視覚的多様性の大小を推定できることを示す．また，推定に有効な画像特徴に基づいて，視覚的多様性と強い関連があるものを調べる．

2. 画像特徴の多様性の分析手法

本研究では，ある形容詞で形容される画像の集合に対して，画像特徴ごとに多様性の強さを示す多様性スコアを定義する．多様性スコアの算出手法の概要を図 2 に示す．具体的には，以下の手順で算出する．

(1) 形容詞 c と関連する，N 枚からなる画像集合 I_c を取得する．

(2) 各画像 i（$1 \leq i \leq N$）から，画像特徴 f の特徴量 $\mathbf{v}_{cfi}$ を抽出する．以降，表記を簡潔にするため，$\mathbf{v}_{cfi}$ を $\mathbf{v}_i$ と表す．

(3) 画像の各組合せ $i, j \in I_c \times I_c$ について，特徴量 $\mathbf{v}_i$,$\mathbf{v}_j$ 間の類似度 $s_f(\mathbf{v}_i, \mathbf{v}_j)$ を計算し，類似度行列 $M_{cf} \in \mathbb{R}^{N \times N}$ を作成する．

(4) M_{cf} を用いて，画像をクラスタリングする．各画像は k_{cf} 個のクラスタ $x_1, x_2, ..., x_{k_{cf}}$ のいずれかに所属する．

(5) クラスタリング結果のエントロピーに基づいて，形容詞 c に対する画像特徴 f の多様性スコア V_{cf} を算出する．

2.1 画像特徴の抽出

形容詞 c と関連する画像集合 I_c（$|I_c| = N$）中の各画像から画像特徴 f を抽出する．本研究は各画像特徴の多様性と視覚的多様性の関係の分析を目的としているため，画像の印象に関わると考えられる様々な画像特徴を抽出する．それぞれの画像特徴 f に対し，2 つの画像の類似度 s_f を定義する．

本研究では，以下の画像特徴 f と類似度 s_f を使用する．

色相（H），彩度（S），明度（V） 入力画像を同一の大きさに拡縮した後，HSV 色空間に変換し，H，S，V それぞれについてヒストグラムを作成する．類似度として負の Earth Mover's Distance（EMD） [9] を使用する．輸送コストには，ビン同士の距離（隣り合ったビンの距離を 1 とする）を用いる．ただし，H に関しては，両端のビン同士が隣り合っているものとして扱う．

局所特徴（AKAZE） Accelerated-KAZE（AKAZE） [7] を用いて特徴点を検出し，Bag-of-Features（BoF） モデル [1] を適用して作成したベクトルを，主成分分析（PCA） [3] により次元削減して特徴ベクトルを作成する．類似度として，重み $\frac{1}{\|\mathbf{v}_1+\mathbf{v}_2\|+\varepsilon}$ を乗じた負の Euclidean 距離を使用する．

シーン特徴（GIST） GIST 特徴量 [6] を使用し，PCA により次元削減して特徴ベクトルを作成する．類似度として負の Euclidean 距離を使用する．

高レベル特徴（CLIP） Contrastive Language–Image Pre-training（CLIP） [8] の画像エンコーダでエンコードされた特徴ベクトルを，PCA により次元削減して特徴ベクトルを作成する．類似度として負の Euclidean 距離を使用する．

なお，PCA による次元削減は，それぞれの画像特徴について，全ての形容詞から取得した画像の特徴ベクトルの

集合に対して行なう．

2.2 多様性スコアの算出

画像集合 I_c 中の全ての画像から抽出した画像特徴 f に基づいて，全ての画像の組合せ $i,j \in I_c \times I_c$ について類似度 $s_f(\mathbf{v}_i, \mathbf{v}_j)$ を計算することによって，類似度行列 $M_{cf} \in \mathbb{R}^{N \times N}$ を作成する．各類似度行列は，画像特徴ごとに全ての形容詞間での最小値が 0，最大値が 1 になるように正規化する．

その後，形容詞 c に対する類似度行列 M_{cf} を用いて画像をクラスタリングし，その結果に基づいて画像特徴 f の多様性スコア V_{cf} を算出する．

本研究では，クラスタ数を事前に定めず，類似度行列に基づいて画像をクラスタリングする手法として，アフィニティ伝播法 [2] を用いる．画像特徴量の類似度行列 M_{cf} に基づいてクラスタリングすると，各画像は k_{cf} 個のクラスタ $x_1, x_2, ..., x_{k_{cf}}$ のいずれかに所属する．

次に，クラスタリング結果に基づいて，画像集合内で画像特徴のばらつきが大きいほど高いスコアをとるように画像特徴の多様性スコアを算出する．本研究では，Jeong ら [4] が提案した視覚性算出手法を参考に画像特徴の多様性スコアを定義する．入力した形容詞に基づく画像集合をクラスタリングした結果のエントロピーに対し，クラスタ内での画像特徴の類似度の低さを表すクラスタ内分散と，クラスタ間での画像特徴の類似度の低さを表すクラスタ間分散によって重みづけする．

形容詞 c に対する，画像特徴 f に関する，クラスタ x のクラスタ内分散 P_{cfx} を以下のように定義する：

$$P_{cfx} = \frac{1}{|x|} \sum_{v_i \in x} \{1 - s_f(\mathbf{v}_i, \overline{\mathbf{x}})\} \tag{1}$$

ここで，$|x|$ は，クラスタ x の大きさを，$\overline{\mathbf{x}}$ はクラスタリング時に算出されたクラスタ x の代表画像の画像特徴を表す．

形容詞 c に対する，画像特徴 f に関する，全体のクラスタ間分散 Q_{cf} を，以下のように定義する：

$$Q_{cf} = \frac{\sum_{i=1}^{kcf} \sum_{j=1}^{kcf} \{1 - s_f(\overline{\mathbf{x_i}}, \overline{\mathbf{x_j}})\}}{k_{cf}(k_{cf} - 1)} \tag{2}$$

エントロピーとして，画像が特定のクラスタ x_i に所属する事前確率 $p(x_i) = \frac{|x_i|}{N}$ に基づく Shannon エントロピー $-\sum_{i=1}^{k_{cf}} p(x_i) \log p(x_i)$ を用いる．これに P_{cfx}, Q_{cf} による重みづけを行ない，形容詞 c に対する，画像特徴 f に関する画像特徴の多様性スコア V_{cf} を，以下のように定義する：

$$V_{cf} = -(1+\beta Q_{cf}) \left\{ \sum_{i=1}^{k_{cf}} p(x_i) \log p(x_i) \left(\alpha P_{cfx_i} + (1-\alpha)\right) \right\} \tag{3}$$

ここで，$\alpha \in [0,1], \beta \in [0,\infty)$ は，それぞれ P_{cfx}, Q_{cf} に対する係数である．

3. 各画像特徴と視覚的多様性の関係の分析

視覚的多様性の推定に有用な画像特徴を明らかにするため，各画像特徴と視覚的多様性の関係を分析した．まず，アンケートによって人間が感じる視覚的多様性を調査するとともに，画像データセットを用いて画像特徴の多様性を定量化した．次に，それらの結果に基づいて，画像特徴の多様性から視覚的多様性の大小を推定した．

3.1 アンケート調査

人間が感じる視覚的多様性を測るために，アンケート調査を行ない，Thurstone の一対比較法 [10] を用いて間隔尺度として視覚的多様性スコアを算出した．Thurstone の一対比較法では試料数が多くなると必要な質問数が大幅に増加するため，本調査では，互いに重複しないように英語の形容詞 10 語で構成される語群を 2 つ用意した．アンケート調査では，2 つの語群のどちらかについて，形容詞 10 語のうち 2 語を全ての組合せで提示し，${}_{10}\mathrm{C}_2 = 45$ 問の質問をした．

アンケートの冒頭では，視覚的多様性に関する説明を提示した．続いて視覚的多様性に関して，2 つの形容詞のみを提示し，それらの単語から画像をイメージして視覚的多様性の大小を比較するよう指示した．アンケート調査には 10 代〜70 代（主に 10 代〜20 代）の 46 人が参加し，その多くは日本語話者であった．各語群で，それぞれ 23 人からの回答が得られた．アンケート調査の回答に，Thurstone の一対比較法を適用して視覚的多様性スコアを算出した結果を表 1 に示す．

3.2 画像特徴の多様性スコア算出

各形容詞に関連する画像を収集するために，Bing 画像検索[*1]で各形容詞をクエリとして画像検索を行ない，検索結果の上位から順に人手で 100 枚の画像を取得した．なお，同じ形容詞 c から取得した画像集合 I_c は常に同一のものとした．

3.1 節でアンケート調査を行なった 20 語の形容詞について，2 章の分析手法を用いて画像特徴の多様性スコアを算出した．このとき，画像特徴として H，S，V を抽出する際には各画像を 256 × 256 画素に拡縮し，ヒストグラムのビン数は 32 とした．また，画像特徴として AKAZE [7] を抽出する際の，次元削減する前のベクトルの次元数は 1,000 次元とし，$\varepsilon = 10^{-10}$ とした．PCA による特徴ベクトルの次元削減では，累積寄与率が 80%を超える最小の次元数を使用した．算出された画像特徴の多様性スコアは，各画像特徴について，全ての形容詞間での最大値と最小値に基づいて正規化した．

[*1] https://www.bing.com/images/ ［2025/5/12 参照］

表 1: アンケート調査に基づいて計算された，形容詞の視覚的多様性スコア．

(a) 語群 1.

形容詞	white	clean	blue	colored	cloudy	cooked	parked	long	old	metallic
視覚的多様性スコア	0.000	0.411	0.234	1.000	0.433	0.808	0.511	0.662	0.725	0.417

(b) 語群 2.

形容詞	black	colorful	yellow	concrete	closed	calm	young	wooden	empty	covered
視覚的多様性スコア	0.000	1.000	0.329	0.328	0.477	0.715	0.812	0.369	0.513	0.707

3.3 推定実験

各画像特徴の多様性スコアについて，視覚的多様性の推定への有効性を調べるため，単語の視覚的多様性を推定するモデルを作成した．表 1 に示した片方の語群で単純なアルゴリズムの機械学習を行ない，もう片方の語群に対する視覚的多様性の推定性能を測定した．

3.2 節で算出した視覚的多様性スコアは，回帰問題として推定すると大きな誤差が生じる可能性がある．これは，語群 1 と語群 2 における視覚的多様性スコアの関係が不明なためである．そのため，本実験では，スコアを直接推定する代わりに，2 つの単語のうち視覚的多様性が大きい方を推定する分類問題を考える．具体的には，線形分類器 [11] を使用し，各画像特徴における 2 つの単語の多様性スコアの差を入力とし，視覚的多様性が大きい単語を出力とする 2 クラス分類とした．10 語からなる語群中の全ての語の組合せを使用するため，学習段階とテスト段階のそれぞれでデータ数は ${}_{10}\mathrm{P}_2 = 90$ であった．語群 1 で学習し語群 2 でテストを行なう分類器と，語群 2 で学習し語群 1 でテストを行なう分類器をそれぞれ作成した．

また，6 種類の画像特徴全てを用いた予備実験により，パラメータを決定した．推定性能の指標としては，2 つの分類器のテストにおける F1 スコアの平均を用いた．その結果，線形分類器の正則化パラメータ $C = 0.01$，多様性スコアを算出する際の係数 $\alpha = 0.25, \beta = 0$ のとき最高の平均 F1 スコアが得られた．$\alpha > 0$ のときに推定実験の性能が最良だったため，α を係数とするクラス内分散による重みづけは，視覚的多様性の推定にある程度貢献していると考えられる．一方で，$\beta = 0$ のときに推定実験の性能が最良だったことから，β を係数とするクラス間分散による重みづけはあまり貢献していないと考えられる．

さらに，視覚的多様性をよく推定できる画像特徴の多様性スコアの組合せを分析するため，特徴量選択を行なった．6 種類の画像特徴のうち，1 種類以上の画像特徴の全ての組合せ（$2^6 - 1 = 63$ 通り）について視覚的多様性の推定性能を比較し，性能が最高になる特徴量の組合せを求めた．

以上の結果を表 2 に示す．平均 F1 スコアの最大値は 0.900 であり，画像特徴の多様性スコアを用いて視覚的多様性を高い精度で推定できていることがわかる．また，上位の画像特徴の組合せには H，GIST，CLIP が多く用いられていた．そのため，これらの画像特徴が視覚的多様性により強く関連していると考えられる．

表 2: 分類性能が高かった画像特徴の組合せとその F1 スコア（上位 5 組まで）．✓ はその画像特徴を推定に使用したことを表す．

画像特徴						平均 F1
H	S	V	AKAZE	GIST	CLIP	
✓	✓			✓	✓	0.900
✓	✓		✓	✓	✓	0.878
✓	✓	✓	✓	✓	✓	0.867
✓			✓	✓	✓	0.844
✓				✓	✓	0.844

4. おわりに

本研究では，形容詞に関連する画像から抽出した画像特徴を用いて形容詞の視覚的な多様性を推定する実験を通じて，画像特徴と視覚的な多様性の関係を分析した．推定実験では，特徴選択により，最大で 0.900 の平均 F1 スコアで視覚的な多様性の大小を推定できた．これらの実験結果から，画像の色相やシーン特徴，高レベル特徴は視覚的多様性と強い関連があることが示唆された．

謝辞

本研究の一部は JSPS 科研費 23K24868 の助成を受けた．

参考文献

[1] Csurka, G., Fan, L., Willamowski, J. and Bray, C.: Visual categorization with bags of keypoints, *Proceedings of the ECCV2004 Workshop on Statistical Learning in Computer Vision*, pp. 1–22 (2004).

[2] Frey, B. J. and Dueck, D.: Clustering by passing messages between data points, *Science*, Vol. 315, No. 5814, pp. 972–976 (2007).

[3] Hotelling, H.: Analysis of a complex of statistical variables into principal components, *Journal of Educational*

Psychology, Vol. 24, No. 6, pp. 417–441 (1933).

[4] Jeong, J.-W., Wang, X.-J. and Lee, D.-H.: Towards measuring the visualness of a concept, *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 2415–2418 (2012).

[5] Kastner, M. A., Ide, I., Kawanishi, Y., Hirayama, T., Deguchi, D. and Murase, H.: Estimating the visual variety of concepts by referring to Web popularity, *Multimedia Tools and Applications*, Vol. 78, No. 7, pp. 9463–9488 (2018).

[6] Oliva, A. and Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope, *International Journal of Computer Vision*, Vol. 42, No. 3, pp. 145–175 (2001).

[7] Pablo, A., Jesus, N. and Adrien, B.: Fast explicit diffusion for accelerated features in nonlinear scale spaces, *Procedings of the 24th British Machine Vision Conference*, pp. 13.1–13.11 (2013).

[8] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning transferable visual models from natural language supervision, *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763 (2021).

[9] Rubner, Y., Tomasi, C. and Guibas, L. J.: The Earth Mover's Distance as a metric for image retrieval, *International Journal of Computer Vision*, Vol. 40, No. 2, pp. 99–121 (2000).

[10] Thurstone, L. L.: The method of paired comparisons for social values, *Journal of Abnormal and Social Psychology*, Vol. 21, No. 4, pp. 384–400 (1927).

[11] Vapnik, V. N.: Pattern recognition using generalized portrait method, *Automation and Remote Control*, Vol. 24, No. 6, pp. 774–780 (1963).

[12] Yanai, K. and Barnard, K.: Image region entropy: A measure of "visualness" of Web images associated with one concept, *Proceedings of the 13th ACM International Conference on Multimedia*, pp. 419–422 (2005).