Vol.2018-CVIM-211 No.4
2018/3/1

Web画像の分布に基づく単語概念の視覚的な多様性の推定

カストナー マークアウレル[1,a)]　井手 一郎[1,b)]　川西 康友[1,c)]　平山 高嗣[2,d)]　出口 大輔[3,e)]
村瀬 洋[1,f)]

概要：近年，画像処理の発展に伴い，画像と自然言語を結び付ける技術が重要になっている．一般に概念辞書のようなコーパスは視覚性や画像特徴を考慮していないため，画像から自然な説明文を自動生成する際や機械翻訳における適切な語彙選択などの障害になっている．そこで本報告では，単語概念の視覚的多様性を推定する手法を提案する．提案手法では，まず，既存の画像コーパスに対して，Web 画像の分布に基づいて決定した重みを利用して，理想的な分布に近づくように各概念に含まれる画像を再構成する．そして Mean-Shift 法による画像特徴のクラスタリングの結果得られるクラスタ数から視覚的多様性を推定する．クラウドソーシングを用いた被験者実験によって決定した真値を用いた評価実験により，15 語の名詞について既存の画像コーパスをそのまま用いた場合よりも正確に視覚性多様性を推定できることを確認した．

Marc A. Kastner[1,a)]　Ichiro Ide[1,b)]　Yasutomo Kawanishi[1,c)]　Takatsugu Hirayama[2,d)]
Daisuke Deguchi[3,e)]　Hiroshi Murase[1,f)]

1. はじめに

近年，Web 画像や SNS の普及とともに，画像と自然言語を関連づけて処理する手法が必要になりつつある．しかし，画像と自然言語の間には，いわゆるセマンティックギャップ（Semantic Gap）と呼ばれる問題が立ち塞がり，計算機による説明文の自動生成や機械翻訳などで障害になっている．画像内容を表現する際に自然な語句を自動的に選び出すためには，画像と自然言語を結び付ける技術が必要になる．例えば，「乗り物」というような抽象的な概念の中には様々なものが存在するために，単語概念と画像特徴を結びつけることは困難である．一方，特定の型式の自動車など極めて具体的な概念は結びつけやすい．

本報告では，このような単語概念の視覚的多様性を推定する手法について検討する．図 1 に従来の単語概念間の距離と，提案する視覚的多様性の違いを示す．図 1(a) は，2 つの単語概念間の距離を計算し，それを視覚的な相違として表した例である．これに対して，図 1(b) の視覚的多様性の場合は，個々の単語概念内における視覚的多様性の度合の推定を目的とする．

視覚的多様性の度合を推定する際に，既存の概念別画像コーパスをそのまま使うと，画像の分布の偏りが問題になる．そのため，Web 上の画像の分布に基づいて，既存の画像コーパスを再構成し，理想的な分布に近づける手法を提案する．

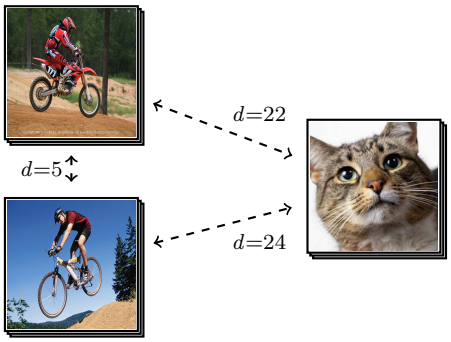
以降，2 節で関連研究を紹介し，次に 3 節で視覚的な多様性の推定手法を提案する．4 節では画像コーパスの再構成により測定した重み，被験者実験により決定した真値について紹介したうえで，5 節で提案手法により再構成した画像コーパスと既存の画像コーパスの各々を用いて評価した実験とその結果を示し，最後に 6 節で本報告をむすぶ．

2. 関連研究

本節では，セマンティックギャップ及び単語概念に関連する研究を紹介する．

情報学だけでなく心理学や言語学の分野でも，視覚情報と言語情報を関連づける研究が行われている．Paivio ら [7] は 925 語の名詞について具象性と心像性を比較し，異なる言葉に対する文字通りの意味とニュアンスの差異を分析し

[1] 名古屋大学大学院情報学研究科
Nagoya University, Graduate School of Informatics
[2] 名古屋大学未来社会創造機構
Nagoya University, Institute of Innovation for Future Society
[3] 名古屋大学情報戦略室
Nagoya University, Information Strategy Office
[a)] kastnerm@murase.is.i.nagoya-u.ac.jp
[b)] ide@i.nagoya-u.ac.jp
[c)] kawanishi@i.nagoya-u.ac.jp
[d)] takatsugu.hirayama@nagoya-u.jp
[e)] ddeguchi@nagoya-u.jp
[f)] murase@i.nagoya-u.ac.jp



(a) 2 つの単語概念間の距離は？

(b) 単語概念の広さは？

図 1: (a) 単語概念間距離の例．2 つの単語概念間の距離として「自転車」は視覚的に「オートバイ」に似ているため距離が小さいが，いずれも「猫」と視覚的に似ていないため，距離が大きくなる．(b) 単語概念の視覚的な多様性の例．左の「スポーツカー」は視覚的な多様性が低いが，右の「乗り物」は多様性が高い．

た．言語学の分野では、単語概念を分類する方法がある．WordNet [4] は英語の概念辞書で，Synset と呼ばれる同義語グループの上位・下位関係を表現した語彙データベースである．秋間ら [10] は Folksonomy を用いて画像特徴とタグを分析し，オントロジーを自動生成する方法を提案した．

一方，視覚的な単語概念についてもいくつかの研究がなされている．最近では，Nakamura ら [6] が複数の画像特徴に対する適応的な重みで概念間距離を測定し，Nagasawa ら [5] が概念間距離に対するノイズ画像の影響を分析した．柳井ら [12] は画像領域エントロピーを分析して，形容詞の視覚性を推定し，その後，小原ら [11] は形容詞と名詞の対について分析した．

Van Leuken ら [9] は視覚的多様化（Visual Diversification）という手法を提案した．視覚的多様化とは，画像検索結果を改良するために，画像特徴のクラスタリングにより，同じ概念の画像を縮退させることである．しかし，視覚的多様性を推定する方法が存在しないため，評価実験として提案手法によるデータセットと人間が作成したデータセットを比較した実験を行った．

3. 視覚的な多様性の推定

従来の概念間距離は 2 つの単語を比較し，相対的な距離を計算したものであるが，1 つの単語の絶対的な視覚性については考慮されていない．視覚的多様化の研究では視覚的多様性を拡充することを目的にしているが，視覚的多様性自体は推定できない．そのため，本研究では，様々な単語概念についてそれ自身の視覚的多様性の度合を推定することを目的とする．

視覚的多様性を推定する場合に，データセットに含まれる画像の分布が問題になる．一般に抽象的な単語は複雑な意味をもつため，様々な下位概念の画像で構成されることが多い．これらの画像の構成比率が多様性推定に影響を与える．WordNet のような概念辞書は視覚的な特性を考慮せずに作成されているため，それらに基づいて構成された画像データセットには偏りがある可能性がある．

そこで，本報告では人々がもつ視覚的多様性を反映した理想的な分布で構成されたデータセットに極力近い分布の画像データセットを再構成するための簡単なアプローチを検討する．各単語概念を用いて理想的な分布のデータセットを構築したうえで，画像から SURF 特徴量を抽出し，Bag-of-Words 手法を利用することで，視覚的なベクトルを作成する．画像の特徴空間内では Mean-Shift 法によってクラスタリングを行い，クラスタ数が単語概念内の視覚的な多様性を示すと考え，クラスタ数を多様性の度合として出力する．

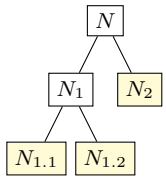
4. 画像コープスの再構築

本節では再構成した画像コーパス及び合成のための重みについて述べる．

4.1 ImageNet 画像コーパスの再構成

ImageNet [1] は WordNet に基づいて分類された 22,000 カテゴリ，1,400 万枚の大規模な概念別画像コーパスである．それぞれの Synset と呼ばれる同義語のグループにはクラウドソーシングによって収集された約 0〜3,000 枚の画像が含まれる．しかし，抽象的な概念は構成画像に偏りがある恐れがあるため，極力，実際の視覚的多様性を反映するように含まれる画像から再構成する．

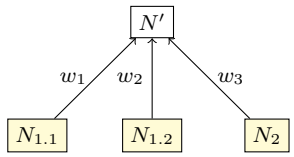
WordNet は，階層化された単語の上位概念・下位概念などの情報を得られる．以降では，グラフ理論の用語を用いて，最下位概念の単語を葉ノード，その上位の単語を親ノードを呼ぶ．親ノードの Synset は下位概念を含むため，より複雑な概念である．まず，1 つの親ノードについてその下位概念をすべて収集する（図 2(a)）．次にその Synset の中で Web 画像の分布に基づく重みを決定する（4.2 参照）．決定した重みは親ノードの概念の中で，下位概念の影響力の大きさを表す．次に，それぞれの下位概念の重みに基づいて ImageNet データセットから画像を収集し，親ノードのデータセットに葉ノードの画像を組み込む．組み込む枚数は重みによって異なるため，下位概念のデータセットの画像が足りない場合には，Web 画像検索 API を用いて，画



(a) WordNet による N の下位概念語の収集.

$N_{1.1} \in N$: w_1
$N_{1.2} \in N$: w_2
$N_2 \in N$: w_3

(b) Web 分布による重みの決定.



N'は
- $N_{1.1}$から w_1枚
- $N_{1.2}$から w_2枚
- N_2から w_3枚

(c) 画像コーパスの構造に従って親ノード画像データセットを再構成.

図 2: 単語概念 N のデータセット再構成の手法.

下位概念	重み
sports_car	27.4%
racer	9.2%
model_t	8.8%
coupe	6.9%
used-car	6.7%
jeep	5.0%
beach_wagon	4.8%
compact	4.5%

(a) Web 分布により決定された car の下位概念の重み.

(b) car のデータセットの再構成．各色は 1 つの下位概念の画像.

図 3: 親ノードに含まれる画像の再構成の例.

像をクローリングする.

4.2 Web 画像の分布に基づく重み

下位概念語の比率に関する分布を計算する際の指標の候補を紹介する．まず，Google API [3] を使用すると，画像検索のヒット数が得られ，これを Web 画像の分布として利用する（図 2(b)）.

葉ノードの最下位概念のヒット数の比率を重みとして，上位概念のデータセットを再構成する（図 2(c)）.

4.3 クラウドソーシングにより決定した真値

視覚的な多様性は主観的なものなので，提案手法の評価実験を行うために，被験者実験によって単語概念の視覚的多様性に関する真値を決定した．ここでは Thurstone の 1 対比比較 [8] を用いた.

被験者実験では「乗り物」のような抽象的な単語から「スポーツカー」や「フォークリフト」のような具体的な単語まで，乗り物に関する 15 個の単語概念について視覚的な多様性を比較した．クラウドソーシングにより，75 名の被験者から 2,139 回の対比較結果を得た．被験者実験の結果に基づき最尤推定（Maximum Likelihood Estimation: MLE）を用いて多様性を表す順位を決定した.

5. 評価実験

本節では提案手法を検証するために行った実験及びその結果と考察について述べる.

5.1 Web 画像の分布で決定した重みで画像コーパス構築

4 節で述べたように，ImageNet を利用した新しい画像コーパスを構築した．図 3 に Web 画像の分布の例を示す．car を上位概念として WordNet からその下位概念語を収集した次に Google API から得られた画像検索のヒット数に基づいて重みを決定し，画像コーパスを再構成した.

Web 画像の分布により，珍しい概念の画像が少なくなって，視覚的多様性の推定に影響を与えにくくなる．一方で sports car のような SNS や広告で人気がある概念は，例えば親ノード car の下位概念としては 27.4%と最大の比率で大きな影響を与える.

表 1: 評価実験の結果.

データセット	順位相関係数	平均2乗誤差
ベースライン手法（ImageNet）	0.45	9.08
比較手法（等比重）	0.73	6.11
提案手法（Web 上の分布に基づく比重）	**0.80**	**4.56**

5.2 既存の画像コーパスと比較

提案手法の有効性を評価するために実験を行った．被験者実験で真値を決定したものと同じ Synset を対象にして，以下の 3 つのデータセットを準備した：

(1) 元の ImageNet 画像コーパス（ベースライン手法）
(2) 重みを等しくして下位概念中の画像を組み込んで再構成した画像コーパス（比較手法）
(3) Google 画像検索により推定した Web 画像の分布（図 3 参照）で決定した重みを用いて，下位概念の画像に偏りをもたせて組み込んで再構成した画像コーパス（提案手法）

各コーパス中の画像から抽出した SURF 特徴量を Bag-of-Words 表現して，特徴空間を Mean-Shift 法によりクラスタリングし，クラスタ数を計算した．正規化されたクラスタ数に基づく順位と Thurstone の 1 対比較法により算出した真値の順位を比較した結果を表 1 に示す．評価指標は，各手法により推定した順位に関する Spearman の順位相関係数（SRC: Spearman Rank Correlation）[2] と，視覚的多様性の値の平均 2 乗誤差（MSE: Mean Squared Error）とした．

元の ImageNet ベースライン手法の結果は被験者実験によって決定された真値とほとんど一致しない．それと比較すると，再構成した画像コーパスは多様性の推定が大幅に改良され，重みを考慮せずに再構成した場合でも相関係数が 62.22%も向上した．提案手法による重みを利用した場合はさらに 9.58%も向上し，真値により近づいた．

6. おわりに

本報告では，単語概念の視覚的な多様性を推定する手法を提案した．また，Web 画像の分布に基づく重みを利用して，理想に近づくように既存の画像コーパスを再構成する手法を提案した．評価実験のために，SNS でのクラウドソーシングによる被験者実験を行い，真値を決定した．評価実験では，本報告で提案した手法により再構成した画像コーパスを用いた場合に順位相関係数が 0.80 となり，ベースラインと比べて 77.78%も向上し，視覚性な多様性をより正確に推定できたことを確認した．

本報告における結果は，今後画像から自然な説明文を自動生成する際や機械翻訳における適切な語彙の選択などの支援になることを期待している．今後の課題としては，WordNet にない単語概念に対応すること，またはメタデータなどのテキスト情報と結びつけた推定などが挙げられる．

謝辞 本研究の一部は，科学研究費補助金及び国立情報学研究所との共同研究による．

参考文献

[1] Deng, J. D. J., Dong, W. D. W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L.: ImageNet: A large-scale hierarchical image database, *Proc. 2009 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 2–9, DOI: 10.1109/CVPR.2009.5206848 (2009).
[2] Dodge, Y.: Spearman rank correlation coefficient, *The Concise Encyclopedia of Statistics*, Springer New York, New York, NY, pp. 502–505, DOI: 10.1007/978-0-387-32833-1_379 (2008).
[3] Google: Google Custom Search API (2016).
[4] Miller, G. A.: WordNet: A lexical database for English, *Comm. ACM*, Vol. 38, No. 11, pp. 39–41, DOI: 10.1145/219717.219748 (1995).
[5] Nagasawa, Y., Nakamura, K., Nitta, N. and Babaguchi, N.: Effect of junk images on inter-concept distance measurement: Positive or negative?, *Advances in Multimedia Modeling: 23rd Int. Conf. on Multimedia Modeling Procs.*, Lecture Notes in Computer Science, Vol. 10133, Springer, pp. 173–184, DOI: 10.1007/978-3-319-51814-5_15 (2017).
[6] Nakamura, K. and Babaguchi, N.: Inter-concept distance measurement with adaptively weighted multiple visual features, *Computer Vision — ACCV 2014 Workshops*, Lecture Notes in Computer Science, Vol. 9010, Springer, pp. 56–70, DOI: 10.1007/978-3-319-16634-6_5 (2015).
[7] Paivio, A., Yuille, J. C. and Madigan, S. A.: Concreteness, imagery, and meaningfulness values for 925 nouns, *J. Exp. Psychol.*, Vol. 76, No. 1, pp. 1–25 (1968).
[8] Thurstone, L. L.: The method of paired comparisons for social values., *J. Abnorm. Psychol*, Vol. 21, No. 4, pp. 384–400 (1927).
[9] van Leuken, R. H., Garcia, L., Olivares, X. and van Zwol, R.: Visual diversification of image search results, *Proc. 18th Int. Conf. on World Wide Web*, pp. 341–350, DOI: 10.1145/1526709.1526756 (2009).
[10] 秋間雄太, 川久保秀敏, 柳井啓司：Folksonomy を用いた画像特徴とタグ共起に基づく画像オントロジーの自動構築, 電子情報通信学会論文誌. D, 情報・システム = The IEICE transactions on information and systems (Japanese edition), Vol. 94, No. 8, pp. 1248–1259 (2011).
[11] 小原侑也, 柳井啓司：Web 上の大量画像を用いた名詞と形容詞の関係分析, 情報処理学会コンピュータビジョン・イメージメディア研究会 (CVIM) (2012/05).
[12] 柳井啓司, コーバスバーナード：一般物体認識のための単語概念の視覚性の分析, 情報処理学会論文誌コンピュータビジョンとイメージメディア（CVIM）, Vol. 48, No. 1, pp. 88–97 (2007).