計画研究　B01-01

画像と言語を用いた質感情報表現のディープラーニング


研究代表者　岡谷　貴之（東北大学大学院情報科学研究科・教授）

研究分担者　川嵜　佳祐（新潟大学大学院医歯学総合研究科・准教授）

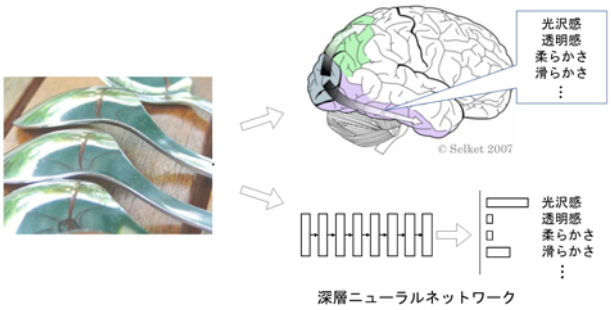

図 1. 目標とする「人と同じように質感を認識できるビジョンシステム」の概念

○研究の背景と目的

われわれは，画像から質感を認識するシステムの実現を目指し，研究を行っています．質感は言語化しづらく，それゆえに人と人，あるいは人と機械の間での共有・伝達が難しいという性質があります．例えば「漆のような黒」や「ダイアモンドの透明感」は，様々に解釈が可能です．これを踏まえ，画像に写る物の質感を「数値化」し，万人が共有できるような「質感の標準化」を可能にすることが，最終的な目標です．

このようなことが実現されれば，産業・社会への大きなインパクトがあります．質感を数値化・可視化することで，工業デザインの効率化（＝製品の試作や実物目視の必要性を排除）を図れるでしょう．また，ウェブの検索サイト等での情報取得や，商品推薦の精度を向上させることもできると期待されます．人が感じとる質感因子の成分が明らかにされることで，画像（に写る物）の質感を制御することや，映像伝送時の情報圧縮への応用も可能になるかもしれません．

質感とは，様々なものを指し得る広範な概念です．われわれは主として，表面の素材や形状に由来する視覚で知覚可能なものを，ターゲットとしていますが，その中にも，光沢感や透明度といった視覚でのみ捉えられるものだけでなく，冷たさや柔らかさといった，触覚に結びついた質感もあります．人はそういった属性であっても，視覚によってかなり正確に認識できることが知られています．

われわれは，こういった質感を，人と同じように認識できるシステムの実現を目指し，いくつかの取り組みを行ってきています．そこで壁になるのは，システムが出力すべきものが人の頭の中にしかない，というこの問題特有の難しさです．また，そもそもどういう概念を扱うべきか，人がどんな画像に，それぞれの概念をどの程度の強さで感じているかなどを測る必要があります．これに付随して，人の知覚にある個人差をどう扱うかも課題となります．このような課題は，物体カテゴリ認識のように，システムが出力すべきものが人と独立に存在する場合には，無縁でした．

以下に，われわれの取り組みのいくつかを紹介します．いずれにおいても，近年，あらゆる画像認識の定番となった畳み込みニューラルネットワーク（以下 CNN）が中心的な役割を果たします．

○これまでに得られた成果

質感形容詞の認識

画像 1 枚からそこに写る物体の（あるいは画像全体から受ける）質感に関わる形容詞の強弱を出力するシステムを構築しました．まず，人の知覚の内容を測るために，2 枚の画像の順序付け（ranking）を人に行ってもらいます．具体的には，2 枚の画像を被験者（クラウドソーシングの作業者）に提示し，複数の形容詞のそれぞれについて，どちらがより強く感じるか（例えば，より“aged”か）と問い，回答を記録します．多数の画像ペアを対象に，複数の形容詞それぞれについて，順序の情報を得ます．

そして，この結果を可能なかぎり忠実に再現するよう，CNN の学習を行います．技術的には以下のようにしています．2 つの入力をとる Siamese 型の構造を持つ，事前学習済みの CNN を使い，画像ペアを 2 つの入力層に入力します．CNN の出力層には，形容詞それぞれ，ソフトマックス関数による出力を持たせ，上述の複数人の順序付けの分布との差を，交差エントロピー損失関数で測り，学習時にはこれを最小化しています．

以上の方法で，各質感形容詞を概ね正確に予測できるようになりました．提案手法の推定結果について，形容詞別に見ると，性能の良い方から順に cold, glossy, wet, aged, transparent, beautiful, resilient, sticky の各形容詞は，われわれが構築したデータセット内では，かなり人に近い認識が出来ています．一方，hard, light, fragile の推定性能は概ね低く，人とはまだ距離があるという結論です．

図 2 にわれわれの CNN の認識結果を可視化したものを示します．これは，各形容詞の強弱の値を計

算した場合に，画像のどこからどれだけの貢献を得たかを表しており，赤が正，青が負の貢献を与えていることを示しています．

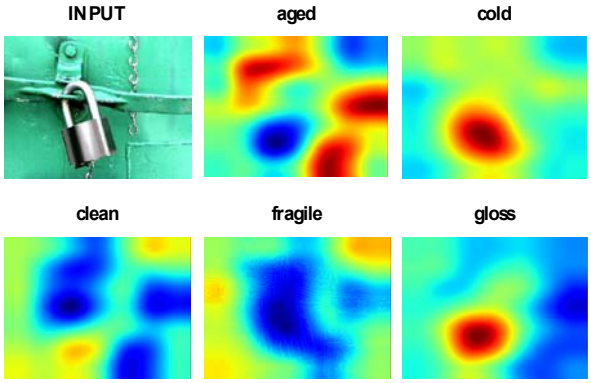

図 2．質感形容詞の CNN による認識の可視化

質感形容詞の発見

上のシステムでは，事前に選んだ形容詞につき，クラウドソーシングによって学習データを作り出していました．しかしこのやり方では，質感の多様な表現すべてを，事前に選び出す必要があり，それは一般に困難であると考えられます．そこで，ウェブ上の何らかのデータを用いて，質感形容詞あるいは類語を，自動的に発見する方法を検討しました．

主として，手作りの商品を販売するサイト Etsy（etsy.com）の商品データを用い，商品画像とその説明文のペアを分析することで，人によるアノテーションを行うことなく，「視覚的に認識可能」な概念を自動的に獲得します．まず，物体認識を学習済みの CNN に，商品画像を入力し，各中間層の出力を記録します．ある特定の語に注目し，その語を説明文に含む商品と含まない商品の２つの集合を考えたとき，前者の画像に対してのみ，選択的に出力が大きくなる中間層のユニットを見つけ出します．このユニットはその語の概念を捉えていると考えることができ，さらに，その語が表す概念は，視覚的に認識可能だと言えます．こうして発見された語と，その可視化の例が図 3 です．

図 3．視覚的に認識可能な概念を表す語彙の発見

質感を含む概念による画像記述

質感の概念の広がりは広く，形容詞一語で表せる質感は限られています．また，質感の表現の中には，文脈によって意味が変化するようなものもあります．読み取った質感を表現するには，形容詞＋名詞でできた句や，より柔軟な言語表現を用いる必要があると言えます．そのためには，最初に紹介した形容詞のように閉じたクラスを対象としていては不可能であり，新しい言語表現を生成できるようにする必要があります．

そこでわれわれは，画像記述の枠組みを採用し，上と同じ Etsy のデータを対象に，商品の画像から，その商品の内容を的確に表現したタイトル文を生成する問題を対象に，この問題に取り組みました．この目的で Etsy の商品のタイトル文から教師データを生成し，これを使って CNN-LSTM ハイブリッドを学習し，画像からタイトル文を生成するシステムを構築しました．結果の一例を図 3 に示します．図には，一般的な画像記述手法の結果（画像とその記述のペアを収めたデータセット MS-COCO を用いて学習したもの）も，併せて示しています．概ね，われわれの方法で，より自然な商品タイトルを生成できることが確かめられました．

図 4．写真からの商品タイトルの生成

○関連する研究発表

論文

1. Vittayakorn S, Umeda T, Murasaki K, Sudo K, Okatani T, Yamaguchi K: Automatic attribute discovery with neural activations. In Proc. European Conference on Computer Vision: 252-268, 2016.
2. Yashima T, Okazaki N, Inui K, Yamaguchi K, Okatani T: Learning to describe E-commerce images from noisy online data, In Proc. Asian Conference on Computer Vision: 85-100, 2016.
3. Li S, Yamaguchi K, Okatani T: Attention to describe products with attributes, In Proc. Machine Vision Applications, 2017.

学会発表など

1. Liu X, Ozay M, Zhang Y, Okatani T: Learning deep representations of objects and materials for material recognition, Vision Sciences Society Annual Conference, 2016.