The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

多様な質感認識の情報処理に用いられる画像特徴を統一的に説明するためのニューラルネットワークモデルの検討

A neural network as a unified model for explaining image features for processing various types of "shitsukan"

上村　卓也[*1]　Takuya Koumura
澤山　正貴[*1]　Masataka Sawayama
西田　眞也[*1]　Shin'ya Nishida

[*1] NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

Natural visual stimuli contain rich "shitsukan", such as glossiness, translucency, and material of an object. Explaining various types of shitsukan in a unified framework is difficult because in general visual perception involves numerous features. Here we tried to explain visual features for shitsukan perception by analyzing the experimental data of shitsukan discrimination tasks. We assumed that participants responded based on the image features of the stimuli. The features were calculated by a deep neural network (DNN) optimized for image classification, in which more complex and abstract features are represented in the higher layers. The features in the middle layer best explained the participants' responses, suggesting that relatively complex features are used for shitsukan perception. We also found that the effective features depends on the type of the shitsukan. These results suggest the effectiveness of a DNN for explaining visual features for shitsukan perception.

1. はじめに

我々は自然な視覚刺激から豊富な質感を知覚できるが、一概に質感と言っても、例えば光沢・半透明性・物体の素材のように、様々な種類が存在する。これまでに、それぞれの質感属性がどのような画像特徴に基づいて知覚されるのかについて、詳細な研究が行われている。しかし、同じ属性であっても質感認識に決定的な画像特徴についての議論は物議を醸している。例えば、物体表面の光沢感については、光沢物体表面の鏡面反射に由来する画像強度ヒストグラムの形状を、視覚系が光沢感推定の手がかりとして用いていることが示唆されている[1]。一方で、同じヒストグラムを持つ物体画像であっても、画像内で鏡面反射によるハイライトとシェーディングとの空間的な位置が整合しない場合には、ハイライトは白い塗装として知覚されて光沢が感じられない[2]。このことは、ハイライトの弁別にはより高次の画像特徴も必要とされることを示唆する。また、半透明物質は、エッジが曖昧で非鏡面反射成分のコントラストが弱くなる傾向があり、それらが半透明性の知覚に利用されることが知られている[3], [4]。画像から切り出した小さな領域のみから物体を構成する素材をある程度推定可能であることを示した computer vision の研究もある[5] 。しかしながら、半透明性や素材についての情報が階層的な視覚情報処理過程のどの段階で捉えられているかについては未だ不明瞭な点が多い。

自然画像の知覚に用いられる特徴を計算するモデルとして最も成功しているものは、深層ニューラルネットワーク（DNN）である。DNN は単純な非線形演算の層を縦列させた構造をしており、特に各層で畳み込み計算を行う DNN がよく用いられ、畳み込みニューラルネットワークと呼ばれている[6]。DNN は膨大な数のパラメータを含んでいるため1つのモデルで膨大な数の特徴を表現することができる。これらのパラメータを自然画像の分類に最適化することによって、高精度な画像分類が可能となる[7]。さらに、画像分類への最適化によって獲得された表現は、動物の視覚系における情報表現とよく似ている[8], [9]。つまり、画像分類に最適化された DNN では自然画像を効果的に表現する特徴が計算され、その特徴は視覚系によって計算されるものと類似している。計算される特徴は、層が深くなるにつれて徐々に複雑で抽象的になる。入力に近い下層では色や傾きのような単純な特徴が表現され、出力に近い上層ではより複雑な形状や物体の種類・概念に関する抽象的な特徴が表現されている[10], [11]。

本研究では、多様な画像特徴が多様な質感知覚を引き起こすという、画像特徴と知覚される質感との複雑な多対多の関係を統一的に理解することを目的とし、そのための第一段階として、様々な質感の知覚に用いられる画像特徴を DNN の層によって表現することを試みた。知覚に用いられる画像特徴を質感の属性ごとに調べるために、ヒトによる質感弁別課題のデータセットを利用した(Sawayama, Dobashi, Okabe, Hosokawa, Saarela, Olkkonen, & Nishida, in prep.)。このデータセットには、質感弁別課題の刺激として用いた画像と、実験参加者による回答が含まれている。課題は、光沢の強さの弁別・光沢の鋭さの弁別・半透明性の弁別・金と黄色プラスチックの弁別・銀とガラスの弁別・光沢と白い塗装の弁別、の 6 種類の質感弁別課題から構成されており、4 枚の画像の中から見た目の異なる画像 1 枚を選ぶという仲間外れ検出課題である（図 1）。本研究では、画像特徴が他と異なっている画像ほど仲間外れとして選ばれやすい、という作業仮説の元、画像特徴を入力とし、課題における各画像の選択されやすさを出力するモデルを仮定した。作業仮説の妥当性を仮定すると、質感知覚に重要な画像特徴からは、実験データとの相関の高い出力が得られると予想できる。画像特徴の候補として、画像分類に最適化された DNN の各層の値を検討した。

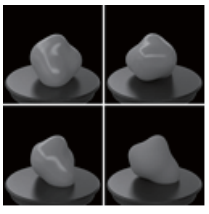

図 1｜実験で提示された画像の例。この例では、右下の画像の光沢の強さが他の画像と異なっている。

連絡先：上村卓也，NTT コミュニケーション科学基礎研究所，
〒 243-0198 神奈川県厚木市森の里若宮 3-1，
koumura@cycentum.com

2. 方法

2.1 刺激画像

データセットには、6 種類の質感弁別課題について、それぞれの次元の質感が異なる複数の画像が含まれている。ここでは画像の合成方法について簡単に説明する。画像の描画には Mitsuba レンダラが用いられている[12]。

光沢の強さと鋭さは、Pellacini et al.による材質の双方向反射率分布関数（BRDF）のモデルにより制御されている[13]。このモデルには光沢の強さを決定するパラメータ c と鋭さを決定するパラメータ d がある。光沢の強さの弁別課題には、d を 0.94 に固定し、c を 0.00 から 0.12 まで 0.02 刻みで変化させた画像が用いられている。光沢の鋭さの弁別課題には、c を 0.06 に固定し、d を 0.88 から 1.00 まで 0.02 刻みで変化させた画像が用いられている。

半透明の物体は、材質による光の吸収・散乱係数として Jensen et al.によって計測された「Wholemilk」のパラメータを用いて、等方性位相関数を用いて描画されている[14]。物質の半透明性は、半透明物質の密度に関するパラメータによって制御され、パラメータの値は 0.0039、0.0156、0.0625、0.25、1.00 である。物体の表面は滑らかな誘電体の素材であるとして描画されている。

金と黄色プラスチックの弁別課題には、金とプラスチックの BRDF を線形補間した画像が用いられており、金の比率が 0.0 から 0.8 まで 0.2 刻みで変化している。銀とガラスの弁別課題の刺激も、銀と誘電体ガラスの BRDF の線形補間によって合成され、銀の比率が 0.0 から 0.8 まで 0.2 刻みで変化している。

白い塗装を施された物体の画像は、光沢物体の鏡面反射によるハイライトが拡散反射によるシェーディングと空間的に整合しない位置に配置されることで生成されている[2]。具体的には、まず、上述の光沢課題と同様のモデルで鏡面反射成分のみを持つ物体画像が描画される。次に、描画されたハイライト画像が拡散反射成分を持つテクスチャとして物体の任意の位置にマッピングされ、再度その物体が拡散反射物体として描画される。生成された白塗装画像は対応する光沢画像とヒストグラムマッチングによって色分布を同一に揃えられている。光沢のパラメータは c = 0.08 とし、d を 0.88 から 1.00 まで 0.04 刻みで変化させた画像が用いられている。

データセットには、それぞれの質感パラメータについて 5 種類の形状の物体を等間隔に 5 方向から見た画像が含まれている。画像の枚数は合わせて、36 種類の質感パラメータ×5 種類の形状×5 方向の視点 = 900 枚である。照明には Bernhard Vogl によるデータベースの「Overcast Day/Building Site (Metro Vienna)」が用いられている[15]。元のデータセットには他の照明下で描画した画像も含まれているが、本研究では 1 種類の照明下での画像のみを用いた。

2.2 実験

データセットには、416 人の実験参加者による回答が含まれている。実験はクラウドソーシングサービスを介して行われたため、実験参加者は各自のコンピュータやタブレット端末を用いて実験に参加している。

実験の各試行では、4 枚の画像が同時に提示され、他と異なって見える画像を 4 枚の中から 1 枚選ぶように実験参加者に指示をしている。4 枚の画像は、全て同じ形状の物体で、異なる視点の画像である。4 枚中 3 枚が同じ質感パラメータを用いて合成された画像で、残り 1 枚が他の画像とは異なる質感パラメータによる画像である。4 枚の画像の配置はランダム、弁別の対象となる質感の属性と物体形状も試行ごとにランダムである。3517 通りの 4 枚の画像の組（順不同）について、各 4 枚組が 20.8 ± 4.5 人の実験参加者に提示された（平均±標準偏差）。

2.3 モデル

本研究では、質感の知覚に用いられる画像特徴を調べるために、画像特徴を入力とし、それぞれの画像が仲間外れとしてどの程度選ばれやすいかを計算するモデルを仮定した。ある画像特徴をモデルに入力して計算された出力が、実験で仲間外れとして選ばれた割合と良く相関した場合、その特徴は質感知覚に重要である可能性が高い。モデルでは、画像から計算された特徴量空間において他の画像から離れた画像ほど仲間外れとして選択されやすい、と仮定した（図 2）。まず、画像 x から関数 f によって特徴量ベクトル f_x が計算される。

$f_x = f(x)$.

f_x と f_y との距離を d_{xy} とし、画像 x から他の 3 枚の画像への距離の平均を d_x とする。

$d_x = \|f_x - f_y\|$,

$$d_x = \frac{1}{n-1}\sum_{y!=x} d_{xy}.$$

ただし n は一度に提示される画像の枚数であり、本研究では n = 4 である。距離 d はユークリッド距離であるとした。画像 x が仲間外れとして選ばれる確率 r_x が d_x に比例すると仮定すると、

$$r_x = \frac{d_x}{\sum_{x'} d_{x'}}.$$

ここで、r_x は、合計が 1 となるように d_x を合計値で割った値とした。

実験で画像 x を仲間外れとして選んだ実験参加者の割合を t_x とし、4 枚×3517 組の画像に対する r_x と t_x の Spearman 順位相関係数を計算した。

このモデルを作業仮説として仮定すると、モデルの出力と実験データとの相関が高くなるような特徴量は、仲間外れ画像を選ぶ際に重要な特徴であると考えることができる。

相関の有意性を判断するために、実験参加者がランダムに回答したと仮定した場合の相関係数を計算した。それぞれの画像の組内で t_x をランダムに入れ替えて r_x との相関係数を 2000 通り計算した。元のデータを用いて計算した相関係数がその中の 99 百分位数より大きいとき、相関が有意であると定義した。

2.4 画像特徴量

質感知覚に用いられる特徴量の候補として、画像分類に最適化された DNN の各層の値を検討した。画像分類に最適化された DNN では層が深くなるにつれて画像から物体カテゴリに

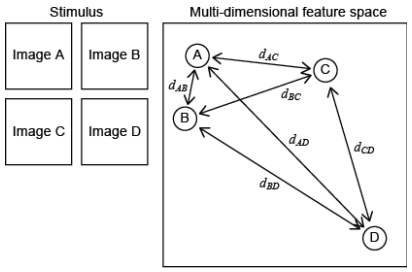

図 2｜画像特徴から仲間外れとしての選ばれやすさを計算するモデルの模式図。4 枚の画像が刺激として提示されたとき、各画像が特徴量に変換され、特徴量空間において各画像間の距離が計算される。他の画像からの平均距離が大きいほど、仲間外れとして選ばれやすいと仮定した。図の例では、画像 D が最も選ばれやすく、次に画像 C が選ばれやすい。

変換される間の抽象化のレベルが高くなる。具体的には、ILSVRC 2012 のデータセットを用いて学習された VGG19 を用いた[7]。VGG19 は、畳み込み演算と半波整流活性化関数からなる畳み込み層が 16 層と、その上の 3 層の全結合層から構成されている。畳み込み層の間には 5 層の pooling 層が存在する。全結合層の活性化関数は、始め 2 層が半波整流関数で、最後の 1 層が softmax 関数である。本研究では、16 層の畳み込み層、5 層の pooling 層、3 層の全結合層の値を画像特徴量として用いた。畳み込み層と全結合層では、活性化関数適用後の値を用いた。最上層のみ、softmax 関数適用前の値も検討した。合わせて、16 + 5 + 3 + 1 = 25 層による特徴を検討した。比較のために、全層の特徴を結合してひとつの特徴としたものも検討した。

比較対象として、テクスチャ画像を表現するために考案されたテクスチャ統計量と[16]、画像の RGB 値も検討した。DNN の値と RGB 値は空間位置で平均した。テクスチャ統計量は元々空間位置で周辺化されているため、そのままの値を用いた。

3. 結果

3.1 モデルの出力と実験データとの相関

DNN の各層の値・全層の値・テクスチャ統計量・RGB 値のそれぞれの特徴量について、モデルによって計算された画像の選ばれやすさと、実験において選ばれた割合との相関係数を計算した（図 3）。相関係数が最も大きくなった画像特徴は、DNN の 25 層中 11 層目の値であった（図4、Spearman 順位相関係数 = 0.59）。11 層目における相関係数は、全層による相関係数（0.58）より僅かに大きく、テクスチャ統計量による相関係数（0.49）や RGB 値による相関係数（0.36）よりも大きかった（図 3）。ただし、本研究の方法では、相関係数の差の絶対値がどの程度大きいと意味のある差といえるのかはわからない。

検討した全ての特徴が、実験参加者による選択の割合をランダムに入れ替えた場合の相関係数の 99 百分位数よりも大きかった。よって、検討したどの特徴を用いても、ヒトの行動データと有意に相関する出力を得られたといえる。

3.2 質感の属性ごとの検討

次に、質感の属性ごとにデータセットを分割し、知覚に用いられる特徴を検討した。質感の属性ごとに、モデルの出力と実験で選ばれた割合との相関係数を計算した（図 5）。実験では、6 種類の質感の弁別課題が行われている。相関係数が最大となる DNN の層が質感の属性によって異なっており、6 種類の質感属性が 3 種類に分類できるようにみえる。黄色と金色プラスチックの弁別と、銀色とガラスの弁別課題では、相関が高い層が 1 層目から約 12 層目まで幅広く広がっていた。光沢感の強度と鋭さについては、6 層目付近で最も相関が高かった。そして、半透明性の弁別と光沢と塗装の弁別では、12 層目付近で最も相関が高かった。相関が高くなる層が深いほど、その質感の知覚には複雑で抽象的な特徴が用いられることが示唆される。

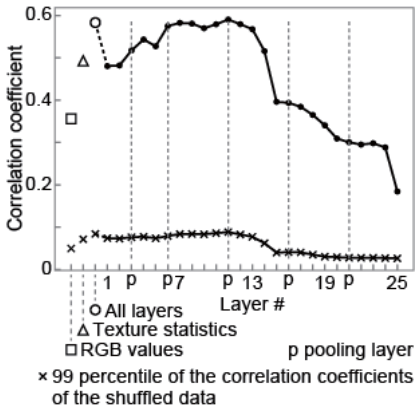

図 3｜各特徴量によるモデルの出力と実験データとの相関係数。横軸に特徴の種類（DNN の各層・全層・テクスチャ統計量・RGB 値）を示した。横軸の p と破線は pooling 層を表す。X 字は実験参加者による選択の割合をランダムに入れ替えた場合の相関係数の 99 百分位数を表す。

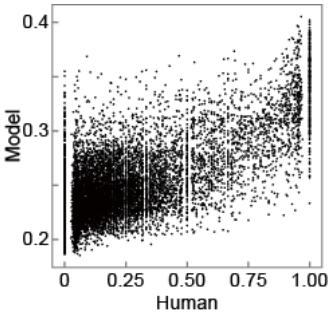

図 4｜DNN の第 11 層の値をモデルに入力したときの出力（縦軸）と、実験において各画像が選ばれた割合（横軸）。1 点が 1 枚の刺激画像を表す。

相関係数の値は質感の属性や DNN の層により大きくばらついていたが、光沢の鋭さ以外の全ての質感属性において、検討した全ての特徴による相関係数が、実験参加者による選択の割合をランダムに入れ替えた場合の相関係数の 99 百分位数よりも大きかった。光沢の鋭さの弁別課題においても、多くの層の相関係数が、ランダムに入れ替えた場合の相関係数の 99 百分位数よりも大きかった。よって、モデルに入力する特徴を適切に選べば、本研究で検討した全ての質感の属性において、モデルの出力が実験データと有意に相関したといえる。

光沢の鋭さの弁別課題では全ての特徴において相関係数が他の課題より小さく、光沢と塗装の弁別課題では層ごとに相関係数が大きく異なっていた。半透明性の弁別課題においてテクスチャ統計量による相関係数が DNN による相関係数よりも大きかったが、他の課題では、DNN による相関係数の最大値が、テクスチャ統計量や RGB 値による相関係数よりも大きかった。

4. 結論・考察

様々な質感を持った画像を刺激として用いた仲間外れ検出課題の実験データを分析した。検討したすべての質感の属性において、モデルによって推定された各画像の選ばれやすさと実験おいてその画像が選ばれた割合とが、有意に相関していた。よって、仲間外れ検出課題の回答を、画像の特徴量空間における距離によってある程度説明できたといえる。特徴の複雑さについて検討するために、特徴量空間を、DNN の各層の出力値によって構成した。ヒトの反応と最も近いパターンを示す特徴は、25 層中 11 層目であった。画像分類に最適化された DNN では、層が深くなるほど複雑で抽象的な特徴が表現される。この結果から、質感の知覚にはある程度高次な情報処理が必要であることが示唆される。

質感の種類ごとに同様の分析を行った結果、黄色と金色プラスチック、銀色とガラスの弁別課題では下層から中間の層における特徴を用いると実験結果とよく相関することがわかった。DNN の下層では、色や傾きの情報が計算されると考えられている。層が上がるにつれ複雑な形状を表現可能になるが[10], [11]、そこには下層で抽出される情報を含むこともできる。これらの質感の知覚には、下層から中間の層までのどの層でも表現可能な、単純な特徴が用いられていることが示唆される。光沢の強さと鋭さの弁別課題では、6 層目付近の特徴を用いるとヒトの反応パターンと最も相関した。光沢の知覚には少し複雑な特

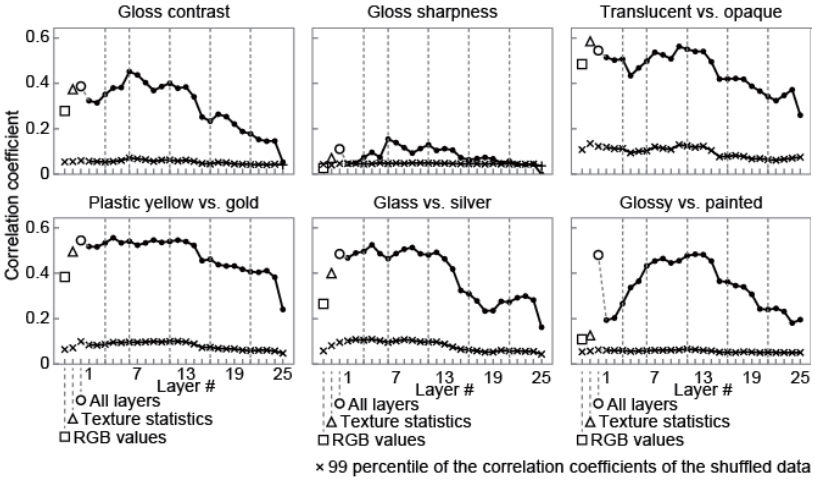

図 5｜質感の種類ごとの、各特徴量によるモデルの出力と実験データとの相関係数。図の形式は図 3 と同じである。

徴が用いられることが示唆される。この結果は、光沢の知覚には画像の輝度分布の高次統計量が用いられるという研究と一貫する[1]。半透明性の弁別課題と光沢と塗装の弁別課題では、12 層目付近の比較的上層の特徴を用いるとヒトの反応パターンと最もよく相関した。これらの質感の知覚には、対象領域の局所的な情報だけでなく、その周囲の情報や物体全体の形状・素材などの情報の統合が必要である可能性が示唆される。

本研究では、モデルの出力と実験データの相関しか調べていない。今後、同様のモデルを用いて未知の実験データを予測できるようになると、より信頼性の高い議論をすることができるだろう。また、今後は DNN の層という抽象的な記述による説明だけでなく、質感知覚と画像特徴の関係をより具体的に解明していきたい。

謝辞

本研究は JSPS 科研費 JP15H05915（新学術領域研究、多元質感知）の助成を受けたものです。

参考文献

[1] I. Motoyoshi, S. Nishida, L. Sharan, and E. H. Adelson, “Image statistics and the perception of surface qualities,” *Nature*, vol. 447, no. 7141, pp. 206–209, May 2007.

[2] P. Marlow, J. Kim, and B. L. Anderson, “The role of brightness and orientation congruence in the perception of surface gloss,” *J. Vis.*, vol. 11, no. 9, pp. 16–16, Aug. 2011.

[3] I. Motoyoshi, “Highlight-shading relationship as a cue for the perception of translucent and transparent materials,” *J. Vis.*, vol. 10, no. 9, pp. 6–6, Sep. 2010.

[4] R. W. Fleming and H. H. Bülthoff, “Low-Level Image Cues in the Perception of Translucent Materials,” *ACM Trans. Appl. Percept.*, vol. 2, no. 3, pp. 346–382, Jul. 2005.

[5] G. Schwartz and K. Nishino, “Perceptual Material Attributes Arise in Local Material Recognition,” *arXiv Prepr.*, Apr. 2016.

[6] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, 1980.

[7] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in *ICLR 2015*, 2015, pp. 1–10.

[8] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, “Performance-optimized hierarchical models predict neural responses in higher visual cortex.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 23, pp. 8619–24, 2014.

[9] S.-M. Khaligh-Razavi and N. Kriegeskorte, “Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation,” *PLOS Comput. Biol.*, vol. 10, no. 11, p. e1003915, Nov. 2014.

[10] A. Mahendran and A. Vedaldi, “Understanding Deep Image Representations by Inverting Them,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[11] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, “Understanding Neural Networks Through Deep Visualization,” 2015.

[12] W. Jakob, “Mitsuba physically based renderer.” 2010.

[13] F. Pellacini, J. A. Ferwerda, and D. P. Greenberg, “Toward a psychophysically-based light reflection model for image synthesis,” in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques - SIGGRAPH ’00*, 2000, pp. 55–64.

[14] H. W. Jensen, S. R. Marschner, M. Levoy, and P. Hanrahan, “A practical model for subsurface light transport,” in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH ’01*, 2001, pp. 511–518.

[15] B. Vogl, “Light probes.” [Online]. Available: http://dativ.at/lightprobes/.

[16] J. Portilla and E. P. Simoncelli, “A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients,” *Int. J. Comput. Vis.*, vol. 40, no. 1, pp. 49–71, 2000.