Vol.2013-HCI-152 No.25
2013/3/14

色，テクスチャ，及びタグ付けされた単語に基づいた画像の印象評価モデルの構築と分析

高橋直己[†1]　加藤俊一[†1]
数藤恭子[†2]　谷口行信[†2]

本研究は，人の持つイメージを可視化することでイメージの共有・伝達を支援するために，視覚刺激とその刺激から受ける印象との関係を，印象の知覚過程に基づいて工学的にモデル化することを目的とする．我々はこれまでに風景写真から受ける印象と色・テクスチャを表す特徴量との関係を機械学習によってモデル化する研究を進めてきた．視覚の感性モデルでは，人間が視覚刺激を受けてからその刺激に対する印象を感じるまでに，色・テクスチャの認識だけでなく，写っている物体（オブジェクト）を認識し，それらの情報を総合的に利用していると考える．本研究ではこの考え方に基づき，オブジェクトを表す特徴量として，画像にタグ付けされた単語を導入し，色・テクスチャの特徴量と併せて用いた．実験では，人物写真のデータセットから抽出された色，テクスチャの特徴量，及びタグ付けされた単語の特徴量と，印象を表すイメージ語との関係を，分析・再利用に都合の良いベイジアンネットワークによってモデル化し，それぞれの特徴量と印象との関係についての分析を行った．

Modeling and Analyzing Impression from Pictures Based on Color, Texture, and Tagged Key Words

NAOKI TAKAHASHI[†1]　TOSHIKAZU KATO[†1]
KYOKO SUDO[†2]　YUKINOBU TANIGUCHI[†2]

In this study, our purpose is modeling based on the perceptual process of impression, the relation between visual stimulus and impressions from the stimulus, in order to support the transmission or shareing personal image by visualizing the image. We have been studying how to model with machine learning the relation between the impression from landscape phoytography and its graphical features that represent color and texuture. In the model of visual perception, we don't recognize only color and texture but also objects in the vision and integrate the information when we are impressed from visual stimulus. In this study, based on the model, we use the tagged key words as graphical features that represent recognition objects with color and texture. In experiment, we modeled the relation between features that represent color, texture and tagged word extracted data set of people photography and image word that represent impression by Bayesian network that is convenient for analyzing or reusing, and analyzed the relation.

1. はじめに

近年のプロダクトデザインにおいては，単にその機能を満たすだけでなく，消費者の感性に響くようなデザインが求められるようになってきている．そのためにメーカーやデザイナーは，消費者がどのようなデザインを求めているかを理解する必要がある．しかしメーカー，デザイナー，消費者のそれぞれのデザインに対する価値観は互いに異なるものであり，メーカーにとっては良いデザインでも，消費者にとっても好ましいデザインであるとは限らない．そのためメーカーは，消費者がどのようなデザインを好むかを調査し，デザイナーと相談してデザインを決めるのに長い時間を費やしている．一方メーカーとデザイナーとの間でも価値観の相違がある．特にデザイナーはデザインに関してのプロフェッショナルであり，クライアントがデザインに関して素人であるときはデザインをまとめるのに非常に苦労する．

このような問題は，価値観や知識の違いによって異なるデザインに対する嗜好を，人々が互いに理解・伝達するのが困難であることに起因する．デザインに対する消費者の嗜好を調べたり，思い浮かべるデザインを他者に伝えたりするとき，私たちは「かわいい」,「高級感がある」などの言葉を用いる．このようなイメージ語は人によって使われ方や解釈がまちまちであり，嗜好やイメージの理解をますます複雑なものにしている．

本研究では嗜好やイメージといった感性情報の相互理解を支援するため，個人の持つ感性を工学的にモデル化し，その特徴を可視化することで，相手の感性の直感的な理解を促すことを目的とする．

2. 研究の背景

価値観や知識の違いによって生じる嗜好やイメージの相違は，個人の感性によるものである．個人ごとの感性の違いは，外界から受け取った刺激を知覚・認知する過程で，それらの情報を個人ごとに異なる経験や知識と関連付けるために生じている．先行研究[1]では視覚感性について，視覚刺激を受けてから感性的な認知が行われるまでの過程を

†1 中央大学
Chuo University
†2 NTT メディアインテリジェンス研究所
NTT Media Intelligence Laboratory

4 つの段階から成るモデルで表現している（図１）.

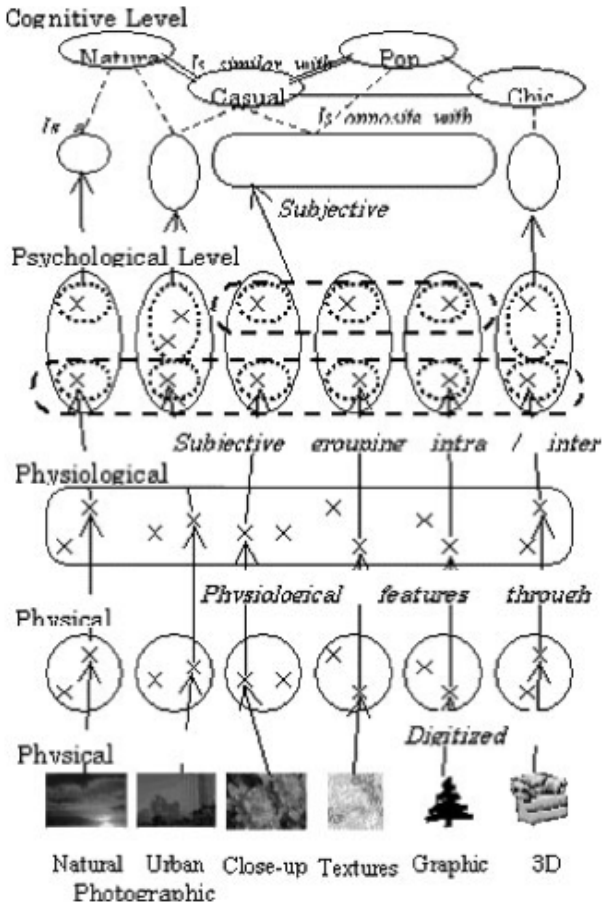


図 1　視覚認知過程の階層的モデル
Figure 1　Hierarchical Modeling of Visual Perception Process.

一方デザインの専門的な知識を持たない人が簡単にデザインを行えるよう支援する研究として，配色やレイアウトなどを提案する方法についての研究がされている[2]．この研究ではカラーポスター制作を支援することを目的とし，ユーザが指定したイメージ語に合ったポスターを作成する．この研究ではポスターのデザイン案を遺伝的アルゴリズムによって作成するため，デザイン生成のアルゴリズムは人間の感性の知覚過程とは異なるものである．そのため，ポスターの制作はできても，その過程からユーザの感性を理解することはできない．

多田ら[3]は，風景写真から受ける印象を個人ごとにモデル化する方法として機械学習を用い，イメージ語を検索キーとする画像検索を行った．この研究では画像から受ける印象を表すイメージ語をアンケートで調べ，画像から抽出される特徴量と対応付けて学習することで印象を自動的に推定する．ここで抽出している特徴量は色とテクスチャを表現するものであり，視覚の感性モデルの生理的レベルでの特徴量に相当する．それに対し，推定したい印象は認知レベルでの情報である．しかし認知レベルでの知覚は，生理的レベルでの知覚をもとに，心理的レベルでの知覚を経てなされるものである．この研究では心理レベルでの知覚に関しては，領域分割によって注目領域を求め，その領域内から特徴を抽出しているものの，領域から抽出する特徴量は色・テクスチャといった生理的レベルのものであり，知識・経験に基づいた情報は考慮されていない．

本研究では視覚の感性モデルを参考に，知識や経験に基づいた特徴を考慮することで，より人間の知覚過程とマッチする印象の推定モデルを構築する．

3. アプローチ

人間の視覚刺激に対する知覚過程に合わせた印象推定モデルを構築するために，生理レベル・心理レベルでの特徴量と，認知レベルでの知覚との関係を機械学習によってモデル化する．3.1 では視覚の感性モデルについて，3.2 では特徴量と学習アルゴリズムについて，3.3 ではシステムの大まかな流れについて説明する．

3.1 視覚の感性モデル

図 1 のモデルでは,物理的な光の信号を受け取ったあと，生理レベルでその情報が色やテクスチャといった情報として知覚される．生理的レベルでの知覚とは人間の初期視覚系における色やコントラストの抽出機構によって知覚される情報である．一方心理的レベルでの知覚とは，網膜から取り入れた大量の情報の中から注目領域の情報だけを選択することであり，特定の領域に注意を向けることに相当する．取捨選択される情報は生理レベルで知覚される色やテクスチャなどの特徴量であり，選択時には明るさの変化が大きい領域を検出したり，自分の知識や経験と照らし合わせたりして注意する領域を決める．そして注目した特定の領域のもつ情報をもとに認知レベルの知覚することで，印象を決定する．

3.2 画像特徴量

本研究では生理的レベルでの特徴量と心理的レベルでの特徴量を両方合わせて用いる．認知的レベルでの知覚には心理的レベルでの知覚（オブジェクトの認識，特定領域に対する注目）だけではなく，生理的特徴量（色やコントラスト）もある程度影響すると考えられるからである．

生理的特徴量は，画像を縦横に 4×4 分割し，各領域の L*a*b のそれぞれの平均値（4×4×3 = 48 次元）と，画像全体の 3 点間コントラスト[3]を 28 個のマスクパターンについて 5 ビンのヒストグラムを求めたもの（28×5 = 140 次元）を用いる．

心理的特徴量としては，知識と経験に基づいた知覚を再現するために，画像内の目立つオブジェクトについての情報を表す特徴量が必要となる．本研究では写真に写っているオブジェクトやその情報をあらわす特徴量として，画像にタグ付けされた単語を用いる．画像にタグ付けをできるサービスは Flickr（http://www.flickr.com/）などの写真共有コミュニティサイトでは一般的なものであり，これらのサ

ービスではタグは検索用のメタデータとして用いられる．これは人がそのタグに関連した画像を検索によって探すということであり，そのタグは画像の特徴的なオブジェクトに関する単語であると考えられる．そこであらかじめ 50 個の単語を用意し，各画像にそれぞれの単語がタグ付けされているかどうかを示す二値の特徴量（50 次元）を用いる．50 単語の選択方法については 3.3 で述べる．

以上計 238 次元の特徴量データを作成し，画像から受ける印象をアンケートによって獲得し，その関係をベイジアンネットワークによって学習する．ベイジアンネットワークはもともと機械学習用のアルゴリズムではないが，統計データを元にしてグラフィカルモデルを作成し，変数にエビデンスを与えて特定のノードの値を確率的に推論することができる[4]．[3]で用いられている SVM は，汎化性能に優れているものの，作成された識別モデルを解析するのが困難という問題がある．ベイジアンネットワークで表現されたモデルはすべてのノードが入力にも出力にもなるので，将来的にはイメージ語を与えることで，そのイメージに近い画像の特徴を予想する機能も期待できる．

3.3 タグ付けする単語の条件

タグつけする単語は膨大な種類があり，それらをすべて特徴量として使うのは無理がある．そこで今回は使用する単語を 50 個に限定する．この 50 個の単語は，画像をイメージ語ごとに分類しやすいよう，以下の値が大きい上位 50 単語を選定した．

$$v_j = V_i\left(\frac{N_{ij}}{N_i}\right)$$

ここでV_iは i に関しての分散を求める関数，N_iはイメージ語 i の与えられた画像の総数，N_{ij}はイメージ語 i が与えられ，かつ単語 j がタグ付けされている画像の総数を表す．ある単語 j が付与されている画像がイメージ語 i である可能性が高い（低い）とき，N_{ij}/N_iの値は大きく（小さく），他のイメージ語 k についてはN_{kj}/N_kの値は小さく（大きく）なりそれらの値の分散が大きくなると考えられる．

なお写真にタグ付けされた単語は，画像内の特徴的なオブジェクトについて表したものであり，それ自体が認知レベルの知覚を表す単語（イメージ語）にならないようにする必要がある．例えば，人物写真には「人」「男性」といった名詞のタグの他，その人の動作などを「走る」「働く」のような動詞のタグで表すことができる．しかし「堂々」「アクティブ」といったその人の状態を修飾語のタグは，タグを付ける人の主観が反映されてしまうため，画像特徴量として不適切である．本研究ではこういった修飾語（形容詞・形容動詞語幹・副詞・「アクティブ」のような英単語としては形容詞になる語）のタグ付けを禁止した．

3.4 システムの概要

ここで本研究での印象提案モデルの大まかなシステムについて説明する（図２）．用意する各画像にはあらかじめ複数のタグが付けられている．これらの画像を被験者に提示し，その画像から受ける印象に最も意味の近いイメージ語を一つ答えてもらう．その後，3.2 で述べたように画像から L*a*b と 3 点間コントラストを画像処理技術によって抽出する．

以上の手順により，各画像に 238 次元の画像特徴量と一つのイメージ語が与えられる．このデータをベイジアンネットワークによってモデル化することで，タグが付いている未知画像のイメージ語を推定することができる．

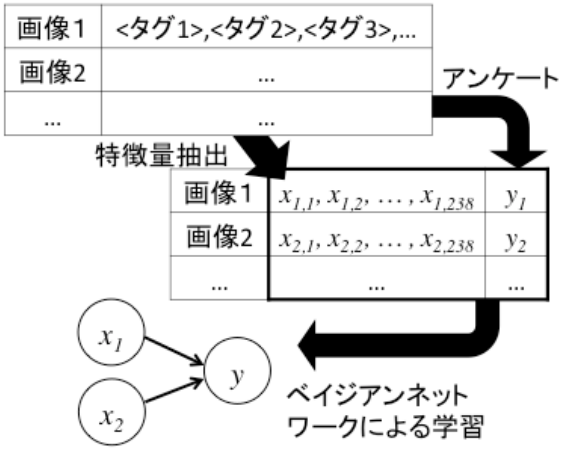


図２　本研究の提案手法のシステム図
Figure 2　System Chart of Described in This Study

4. 実験と考察

提案手法の効果を調べるため，実験を行った．写真はタグが付いた人物写真 1434 枚を用い，被験者にアンケートをとった．イメージ語は，プロの写真家が写真を分類するときに用いる active, classic, cyber, elegant, fresh, natural, modern, pop, pretty, sexy, wild の 11 種類を用いた．タグ付けされた単語は全部で 3232 種類あり，その中から 50 個の単語を 3.3 に述べた基準によって選択した．

4.1 推定の結果

ベイジアンネットワークによる学習では，K2 アルゴリズムによる構造学習を行い，10 fold 交差検定法を用いてイメージ語の推定を行った．表 1 はどの画像がどのようにイメージ語を推定されたかを表している．縦列のイメージ語は正しいラベル，横のイメージ語はシステムによって推定されたイメージ語である．グレーの部分は正しく推定できた回数を表す．

全体としての識別性能は低いものの，ほとんどの場合，グレーの部分の数値が，その列の中で一番大きな値をとっている．一方各行内での最大値は natural に集中している．これより，未知画像の印象を推定するとき，その大部分は natural と推定されているが，推定結果で見ると，各イメージ語グループの特徴を画像特徴量によってある程度抽出，

学習が出来ていると言える．

表１　推定結果ごとの推定の回数
Table 1 The number of times of prediction through each results.

	active	classic	cyber	elegant	fresh	modern	natural	pop	pretty	sexy	wild
active	67	0	1	0	54	10	11	2	6	0	11
classic	3	28	0	7	33	5	30	2	17	0	0
cyber	0	0	37	1	2	9	12	0	0	0	0
elegant	1	0	1	18	41	2	46	8	21	0	1
fresh	16	0	1	0	63	7	54	3	4	0	0
modern	0	0	1	0	15	69	68	0	2	0	0
natural	3	0	0	1	71	7	97	1	10	0	0
pop	2	3	1	10	22	8	59	10	7	0	0
pretty	1	0	1	0	36	0	44	3	34	0	0
sexy	0	0	0	1	12	1	65	3	0	0	1
wild	28	1	0	1	16	2	47	1	1	0	33

4.2 グラフの構造

イメージ語のノードと直接因果関係が認められたのは，単語の変数のみで，色やテクスチャなどの生理的特徴量はグラフ上でイメージ語ノードと矢印が書かれなかった．L*a*b と 3 点間コントラストは，イメージ語ノードとも単語ノードとも関係性は弱く，L*a*b 同士，3 点間コントラスト同士での結びつきが強く評価された．このことから生理的特徴量より心理的特徴量のほうが，イメージ語の表現や識別に適していると考えられ，視覚の感性モデル（図１）との適合性が認められる．しかし，一般的には色やテクスチャのような生理的な特徴も部分的には印象面に影響を与えうると思われる．今回の実験で生理的特徴量とイメージ語間で因果関係が無いと評価されたのは，生理的特徴量の設計方法に問題があったためと考えられる．

本実験で用いた生理的特徴量は，高次元な一つのベクトルで表現され，色またはテクスチャについての情報を保持している．しかしベイジアンネットワークの構造選択の際には，ある単一のノードと別の単一のノードの間に因果関係があるかどうかが判断される．すなわちベクトルのうちの一つの要素とイメージ語の間に関係があるかどうかが構造選択の判断のポイントとなる．しかし今回用いた特徴量はベクトル全体で情報を表現するものであり，構成する要素単体では情報量が少なく，他の要素との強い結びつきがある．そのため，色同士，テクスチャ同士での因果関係が多く評価され，イメージ語との関連性は低く評価されたのであろう．これを解決するためには，色やテクスチャの特徴量を高次元ベクトルのまとまりとして表現するのではなく，互いに独立であり，スカラー量で意味を持つような特徴量を設計する必要がある．方法としては，多数の要素からの統計量を求める方法があり，輝度勾配を扱う研究ではさまざまな統計量が利用されている[5][6].

4.3 精度とデータ数のばらつき

今回の実験では 1434 枚の画像を用いたが, アンケートの結果, それぞれのイメージ語ごとの画像の枚数は, 最大 190（natural），最小 61（cyber）とかなりばらつきが出た．条件付き確率表を調べると，データ不足のために複数のノードにおいて条件付き確率が計算されていないことがわかった．そのため，未知画像の印象推定の際に必要な条件付き確率が分からず，natural と推定してしまっている．

対処法としては（a）教師データを増やす，(b）イメージ語ごとの画像の枚数のばらつきを減らす，(c）特徴量の次元数を縮約して必要データ数を減らすことが考えられるが，（a）は，完全データで学習を行うことを意味するが，現状だとそのために必要になるデータ数は膨大であり，アンケートの実験が困難になる．(b）はイメージ語ごとの画像数はアンケートによって決められるので制御するのが困難になる．画像枚数が少ないイメージ語に合わせて，データ数を減らせば, さらなるデータ不足となってしまう. 一方(c）は，4.2 で述べた特徴量を統計量でスカラー化することで実現する可能性がある．また，ある程度必要データ数が確保出来れば EM アルゴリズムによるデータの補完も方法としてあげられる．

5. まとめ

本研究では視覚の感性モデルをもとにした画像特徴量を利用して，画像から受ける印象のベイジアンネットワークによるモデル化を試みた．実験では満足な識別精度は得られなかったものの，特徴量の設計方法やベイジアンネットワークの学習時の工夫などについて，いくつかの知見を得た．

また実験方法としてアンケートを行う際に，画像から受ける印象を 11 個のあらかじめ決められたイメージ語の中から最も近いものを選んでもらったが，別の方法として，該当するイメージ語を複数選択してもらうことが考えられる．この方法では非常にアンケートにコストがかかるため今回は実施しなかったものの，実際には一つの画像から「pop かつ pretty」というように複数のイメージ語が該当するケースは珍しくない．人によっては，ある二つのイメージ語の意味が近いと感じる場合も有り得, そのような場合，どちらか一つを選びにくく，学習システムもその分類基準を見つけにくくなってしまう．今後としては，4.2 及び 4.3 で述べた特徴量，データ数に関する問題への対処，本章で述べたアンケート方法の検討を進め，より効率よく，効果的な識別ができるシステムの構築を目指す．

謝辞　本研究を進めるにあたり，貴重なご意見を戴く㈱アマナホールディングス，㈱アマナイメージズの皆様，日

頃より，熱心な研究討論や実験への協力を戴く中央大学理工学部ヒューマンメディア研究室の皆様，感性ロボティクス研究センターの皆様に深く感謝します．本研究は一部，科学研究室費補助金・挑戦的萌芽研究”マルチモーダル感性認知機構の言効率なモデル化と実環境快適化への応用”（課題番号 24650110），中央大学理工学研究所・共同研究”感性ロボティクス環境による共生的生活空間の構築と感性サービスへの応用”などの支援を受けて実施しました．

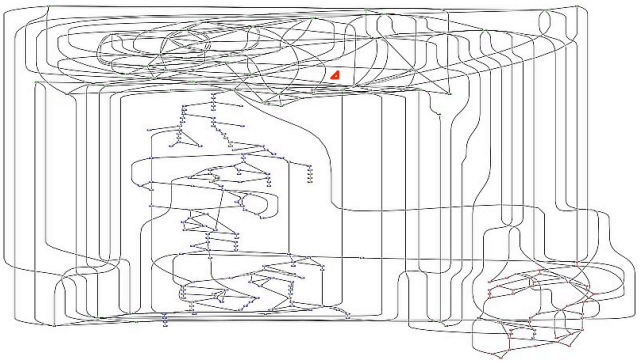
参考文献

1) 加藤俊一: 視覚感性の工学的モデル化とその情報提供サービスへの応用，日本画像学会誌，Vol.47，No.3，pp.183-188 (2008).
2) 尾畑貴信，荻原将文: 感性を反映できるカラーポスター作成支援システム，情報処理学会論文誌，Vol.41，No.3，pp.701-710 (2000).
3) 多田昌史，Zhang Zhongfei(Mark)，加藤俊一: MIL を用いた視覚的印象の分析・学習と画像自動分類への応用，情報処理学会研究会.CVIM，Vol.2006，No.25，pp.13-18 (2006).
4) 本村陽一: ベイジアンネットワーク：入門からヒューマンモデリングへの応用まで，日本行動計量学会第 7 回春のセミナー (2004).
5) Isamu Motoyoshi，Shinya Nishida，Lavanya Sharen，EdwardH. Aselson: Image statistics and the perception of surface qualities，Nature，Vol.447 (2007).
6) Andrea Baraldi，Flavio Parmiggiani: Investigation of the textural characteristics associated with gray level cooccurrence matrix statistical parameters. IEEE Transactions on Geoscienceand Remote Sencing, 33, pp.293-304 (1995).

付録

付録 A.1 ベイジアンネットワークよるグラフ

ノードはイメージ語ノードの他に，単語に関するもの，色に関するもの，テクスチャに関するものの３つに大別される．次の図はベイジアンネットワークの構造選択アルゴリズムによって作成されたグラフであるが，図中でノードがおよそ 3 つの塊になっている．図の上の方にある横に長い塊が単語のノードにおおよそ相当し，図の中央よりやや左の塊はテクスチャと，右下の塊は色とそれぞれ対応している．イメージ語ノードは赤い丸で示した箇所にある．



下表はイメージ語ノードと直接因果関係があると認められたノードである．4.3 に述べたように，単語のノードのみであった．

屋外	スポーツ	日本人	ビジネス
笑顔	サイエンス	レジャー	女の子
日ざし	男の子	未来	元気
ブレ	落ち着き	ジャンプ	行事