人の感覚と人工知能のモダリティ

The survey on the modalities of human and artificial intelligence

柴田 淳司 [1*]
Atsushi Shibata[1*]

[1] 東京都立産業技術大学院大学　Advanced Institute of Industrial Technology
*Corresponding author: Atsushi Shibata, shibata-atsushi@aiit.ac.jp

Abstract　Humans process multiple senses simultaneously, and each sense is called a sensory modality, while the presence of multiple sensory modalities is called multimodality. In recent years, active research has been conducted into the interactions between sensory modalities. In the field of artificial intelligence, the combination of input and output is also called a modality, and research into multimodality has become more active in recent years. In this paper, we introduce examples of research into multimodality between humans and artificial intelligence and consider the benefits that can be obtained from multimodality. If we can elucidate this and objectively judge the harmony that people perceive, this could be applied to a variety of services.

Keywords　sensor modality; multimodal; cross modal; multimodal deep learning; cognitive psychology

1　はじめに

似合う，違和感があるとはどういう状態を指すだろうか．服を買う時，その服が自分に似合うかどうかで買うことがある．あるいは体調が悪い時，喉に違和感があるという表現をすることがある．辞書的定義によると，似合うとは両者が調和していること，違和感があるとは調和を失いチグハグな様子であることとされている．両者に共通するのは，複数のものの調和が取れていないことである．

人が感じる調和は複数の印象の差から生じるものと考えてみる．例えば服とそれを着る人の調和を考える．服には色や質感，意匠が存在する．また，着る人には体格，肌の色など外見に加えて性格などの内面的な特徴が存在する．人がそれらを個別に見た時，服には服の，人には人の印象が湧く．そして，これを組み合わせたとき，両者が持つ印象が衝突することを「違和感がある」と表現し，衝突しないことを「似合う」と呼ぶのではなかろうか．もちろん両者を見た人の文化，経験からくる評価，例えば和服は細身の人の方が似合うなども存在するが，ここでは考慮しない．ではそれらを見た人の印象はどう発生するのだろうか．

人は複数の感覚を同時に処理しており，それぞれの感覚のことを感覚モダリティ，複数の感覚モダリティを持つことをマルチモーダルと呼ぶ．近年では，感覚モダリティ間の相互作用についても盛んに研究がされている．

ところで，人工知能の分野でも，入力と出力の組み合わせをモダリティと呼び，近年はマルチモーダルの研究が盛んとなってきている．

本稿では，人と人工知能のマルチモーダルに関する研究事例を紹介するとともに，マルチモーダルにより得られる恩恵について考察をする．これを解明することにより，人が感じる調和を客観的に判断することができれば，さまざまなサービスに応用が期待できる．

2　人の感覚

感覚と認知

感覚とは，外界の情報を刺激として受け，それにより脳に生じる意識である．古くは紀元前のアリストテレスが五感（視覚・聴覚・嗅覚・触覚・味覚）を提唱しており，それ以降も哲学や医学の分野で，近代では心理学で研究がされている．情報技術の発展により，感覚の研究はさらに派生し，認知科学やAffective Computing などの分野で情報システムとして分析されており，それを活用したデザインやサービスへの応用をされている．

心理学において，感覚とは外界の情報から得られる意識経験のことを指す．その処理の流れを図 1 に示す．外界にはさまざまな物体が存在し，人はその一部を感覚器という器官を用いて捉えることができる．例えば視覚を司る感覚器である眼では，対象から発せられる光が網膜に当たると電気信号に変換され，脳へ送られる．同様に聴覚は耳，嗅覚は鼻，触覚は皮膚，味覚は舌と，感覚と対応する感覚器が存在する．五感と呼ばれるこれらの感覚の他にも，平衡感覚と内耳，熱さと温点など，数多くの感覚と感覚器が人には存在する．

感覚器から得られた信号は受容器を通して電気信号として脳に送られ，これを知覚と呼ぶ．知覚された情報をもとに，人の脳内では思考や言語理解，記憶などの認知が行われる．また，記憶や印象によりフィードバックが起こり，新たな感覚が想起されることもある．

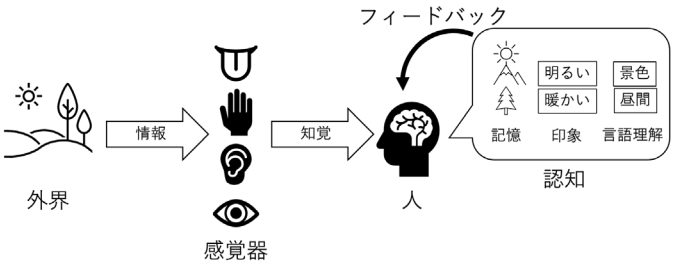

図 1　感覚器を介して外界の情報を知覚し，認知を行う

認知の仕組みは未だわかっていないことが多い．哲学分野では，知覚された外界の情報により脳内に発生する鮮明で明確な質感のことをクオリアと呼ぶ．例えば我々がコップを見たとき，脳内では確かにそれに対応する単語やイメージ，色合いや手触りなどの感覚が呼び起こされる．そしてそのことを外部に伝えるためには画像や声，文字に起こす必要があるため，最終的には容易さの面から言語情報として加工され出力されることが多い．これは機械学習における入力と特徴量，出力の関係に近い．画像物体認識では画像をもとに，その特徴量を計算し，そ

Received: 2024-10-30; Published online: 2025-02-28

こに移っているものが何かを分類する．この時の画像が視覚情報であり，特徴量がクオリア，出力の分類が言語情報と対応する．

以上のことから，人は複数の入力をもとに，特徴抽出を行い，思考するシステムと捉えることができる．この時の感覚と認知の 1 セットの感覚モダリティ（sensory modality）と呼ぶ．モダリティとは手段や様式という意味で，視覚の感覚モダリティなら映像が眼を通して知覚され，認知される一連の流れのことを言う．単一の感覚モダリティであることをユニモーダルと呼ぶ．デジタルカメラは外界をレンズ越しに映像素子で感知し，画像に変換することからユニモーダルと言える．一方，人の場合は視覚だけに限らず，多くの感覚器を備えており，同時に複数の処理を統合しながら行なうため，マルチモーダル（multimodal）と表現される．加えて，複数の感覚が相互に影響しあうことが報告されていることから，クロスモーダル（cross modal）と表現されることも多い．

人が似合う，違和感などは，マルチモーダルな情報を統合した際の齟齬によるものと考えられる．

音から受ける感覚

音は空気の振動であり，鼓膜，耳小骨，蝸牛の有毛細胞の震えにより電気信号に変えられ，脳に送られる．人が認知する音は，対面でのコミュニケーションにおける重要な要素であるが，人に印象を与える大きな要素の 1 つである．ここでは声による印象と，音楽による印象について紹介する．

声による印象を理解するために，まずは声について解説する．人の声は基本となる母音と，それを修飾する子音とに分けられる．母音は肺から出される空気に声帯の振動が加わることで音になり，その口や鼻で共鳴することで作られる．対して，子音は唇や舌の動きにより作られる．日本語における母音は「a」「i」「u」「e」「o」が該当する．これに子音である「k」や「s」を組み合わせることで「ka」や「sa」などの音を作っている．母音が共鳴により作っているため，いくつかの周波数が強い山の部分があり，これをフォルマントと呼ぶ．周波数が低い順に第 1 フォルマント，第 2 フォルマント，第 3 フォルマントと名前がついており，特に第 1 フォルマントの周波数は基本周波数と呼ばれている．人は第 1 フォルマント，第 2 フォルマントの組み合わせにより母音を識別していると言われている．もちろんフォルマントには個人差や話し方による振れ幅があるため，ある程度の範囲内に収まっていれば同じ母音であると認識している．

このように，フォルマントは言語情報を理解するために重要なものだが，印象や話者の感情を読み取るためにも重要な意味合いがある．ここでは，人が感じる印象について取り上げたいので，周波数や長さなど声の物理的情報のみを扱い，言語的な内容については除いて考えてみる．人は言語情報がなくとも，声に含まれる声の高さ，大きさ，長さ，リズム，間など様々な情報を含んでおり，こうした情報を韻律（prosody）と呼ぶ．

まず，一般に女性や子供は喉が小さいため，物理的に出せる音の周波数が高く，男性は低く制限されている．声の印象から，高い方が女性や子供らしく聞こえ，低い方が男性らしく聞こえる．

韻律は感情を伝える手段としても用いられる．日本語話者の音声による感情表現の研究[1, 2]では，喜びを表すときは，平均の基本周波数が高く，レンジが広く，持続時間長は短い傾向がある．また，怒りを表すときは，基本周波数のレンジが広く，持続時間長は短い傾向がある．悲しみを表すときは，平均の基本周波数が低く，レンジが狭く，持続時間長は長い傾向がある．加えて，語尾の基本周波数が上がると不安や疑問を，下がると納得を表す．ただし，これらの感情表現は言語や文化に強く依存するため，必ずしも普遍的なものではないとされている．

音声以外の研究では，音楽を聴いた時の印象の研究[3]がされている．これらの研究によると，長調の音楽は明るい感情を，ゆっくりしたテンポは落ち着きを，速いテンポは興奮を生じやすいとされている．音楽による心理的変化は音楽経験の有無には関係ないという報告[4]がある．

こうした音の性質はマーケティングでも活用されている[5]．スーパーマーケットにおける BGM が，ゆっくりしたテンポだと顧客の滞在時間が長くなり，速いテンポだと顧客の滞在時間が短くなることが示されている．また，テンポが早いと購買行動を促進するが，早すぎると逆効果になることも示唆されている．

光から受ける感覚

人は外界からの情報を素早く認知するため，他の感覚モダリティより優位であるとされている[6]．メラビアン[7]によると異なる情報を同時に与えたとき，その重要度は言語情報，聴覚情報，視覚情報の順に 7%，38%，55%だったという．この結果から視覚は外界の情報を正確に捉えているように感じるかもしれないが，脳内では様々な処理がされ，加工された情報を基に人は認知を行っている．

光は電磁界を介する振動と質量 0 の物質との重ね合わせ，光子と呼ばれる量子によって伝播される．そういう意味では電磁界を伝わる音のようなものである．これは熱量を持ったあらゆる物体から発せられるが，人が捉えられる光子の振動数は波長が 380 nm から 760 nm 程度と範囲が決まっており，これを可視光と呼ぶ．太陽や高温を持つ物体，発光ダイオードや蛍光灯など電子の励起を起こす物体は直接可視光を発するため視覚により認知できるが，それ以外の物体は他の光源からの光子の反射により捉えられる．より厳密にいうならば，人は赤，緑，青色を感じやすい 3 種類の錐体細胞と，少量の光でも反応しやすい桿体細胞，計 4 種類の視細胞の反応によって光を感じている．そのため，実際には虹の色が物理的な色の分布なのだが，人は赤と青の中間に紫を感じるなど，脳内で補完する作用が働いている．また，網膜には視細胞が存在しない箇所が存在するため，本来は視野に盲点と呼ばれる欠けが存在するはずなのだが，脳が補完を行っているため認知されない仕組みとなっている．

形状に関する認知として，解剖学の知見から，視覚マスキングと呼ばれる，特定の形を検出する脳細胞が見つかっている．さらに高次の脳処理では特定の特徴にマッチする専用の脳細胞があるという仮説がある[8]．

一般論として，赤や橙色は暖色，青や紫は寒色，緑などは中間色と表現され，それぞれ暖かい，冷たい，どちらでもない印象があると言われている．また，色の組み合わせにより緑，青，茶色の自然を感じるナチュラルカラーや，黒地に赤や金色のアクセントが高級感を感じるなど，色には様々な印象が付属するとされてきた．

色の印象をアンケート調査し，主成分分析を行なった研究[9]では，単色を見たときに人が感じる印象を，活発さ，重さ，熱さの3項目で表現をしている．例えば，赤は熱い，青は冷たい，黄色は活発で軽い，黒は重い，といった具合である．

虹の色の数が国や地域によって違うように，色の感じ方にも地域差があることが報告[10]されている．大学生を対象としたアンケート調査によると，赤，青，緑，黄色の基本色はポジティブな影響を，それ以外の中間色はネガティブな印象を与えるという結果が報告されているが，文化による差が大きいことも示唆されている．調査では基本色の中でも緑のリラックス効果が低く，その理由を緑が毒を連想させるからだとある．これは欧米では毒は肉が腐敗した色であることに由来する．海外のゾンビ映画でも肌は緑色をしている様子が見受けられる．一方，東アジアの毒は植物由来であり，紫色をしている場合が多く，緑と毒のイメージが結びついておらず，結果に差がついている．

人は色だけではなく，テクスチャからも印象を読み取ることができる．学生を対象に行った実験[11]では，プラスチックや木材などのテクスチャ画像130枚に対して光沢，透明性，色彩，粗さ，硬さなどの項目を正解データとして作成し，他の被験者が画像を見た時に表現した形容詞と比較したところ，高い相関が見られた．このことから，視覚情報だけでも対象の柔らかさなどの情報を脳内で連想させることが示されている．

これら色の印象をマーケティングに利用する研究もされている．店内の色による消費者行動への調査[12]では，暖色系の配色が活気を与えて消費者の興奮や購買意欲を高めるのに対し，冷色系は落ち着きを促し，長く滞在しやすい環境を作り出す効果があると示唆されている．

マルチモーダルな感覚

心理学において，錯覚や誤認に関する研究は古くからされている．その中でも，複数の感覚モダリティを同時に処理することにより生まれる新たな認知や現象がある．

McGurk効果[13, 14]では，被験者に「が（Ga）」と発声している人の映像と，「ば（Ba）」という音を同時に提示すると，「だ（Da）」という音に聞こえるというものである．これは視覚から提示される情報と，聴覚から提示される情報が一致しないため，脳内で情報が競合し，補完した結果，新しい音として認知されるという物である．

視覚と言語情報の錯覚として，色を表す文字と，その文字の色が一致していない場合，色の認知が遅れるストループ効果[15]と，文字の認知が遅れる逆ストループ効果[16]が報告されている．これは，色と文字の形状はどちらも視覚情報だが，脳内では色の情報と言語情報は別の感覚モダリティであり，それらが競合する結果，認知までの時間がかかるというものである．このことから，視覚情報や言語情報は高次の脳処理の過程で統合して処理されることが示唆される．

同様に視覚による印象と言語による印象が同一の印象空間で表現されているという証左として，ブーバ・キキ効果（Bouba/Kiki effect）[17]やタケテ・マルマ効果（Takete/Maluma effect）[18]が挙げられる．図2を見て，どちらがブーバでどちらがキキかを選ぶ，あるいはどちらがタケテでどちらがマルマかを選んでみる．すると，大抵の人は尖った方をキキ，タケテとし，丸い方をブーバ，マルマだと判断する．これはbやmの発音は丸みを帯びた印象を，kやtの発音は尖った印象を与えるため，視覚的特徴と合致したものを選ぶためと考えられている．この効果は音が印象を元に作られたとする音象徴主義の根拠として用いられることがあるが，実際の単語は複数の要因によって決まるため，全てが音の印象と単語の意味が一致しているとは限らない．特に文化的な影響が大きく，英語では光関係の単語の接頭にgl-がつくため，gl-の発音が光の印象と混じっているが，日本人にとっては低い音を発するgは重い印象を受けるため，軽い印象がある光は連想されない．

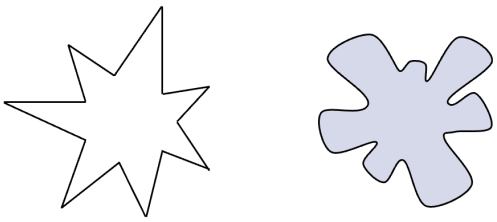

図 2　ブーバとキキ，あるいはタケテとマルマの図

視覚や聴覚，言語認識だけでなく，味覚や嗅覚でもマルチモーダルの影響が確認されている．味と匂いが合致していると，その味を強く感じやすくなり，逆に一致していないと味を感じづらくなるという報告[19]がある．実験ではショ糖にイチゴの匂いを加えると甘いという感覚を加速させ，逆にハムの匂いをつけると甘さを減退させるという結果が示されている．

この味と匂いの相互影響を利用した研究に，MetaCookie+[20]がある．この研究では，プレーンクッキーに匂いをつけることで，仮想的にチョコやナッツなどの味を想起させるというものである．

こうした複数の研究結果より，人は複数の感情モダリティを統合することで，認知を加速させたり，減衰させたり，誤認するということがわかった．この性質をうまく利用すれば，MetaCookie+のように，ポジティブな認知を想起させるサービス設計へ応用できる可能性がある．例えば，ロボットには不気味の谷という概念がある．

ロボットには不気味の谷現象[21]があるとされている．人に似せたロボットを作った際，似せるほど違和感を感じるという現象である．この現象もマルチモーダルの観点から考えると，人に似た外見だが，人とは違う振る舞いによる印象の誤差が，特に日常生活で識別すべき「人」と「それ以外」に対して厳密に働き，不気味さとして認知されるのかもしれない．最近では，こうした感覚間で相互作用を及ぼすものをクロスモーダルと呼び，研究[22]されている．例えばUI設計において，視覚的な情報と触覚的な操作の相互作用を考慮するなどがされている．

加えて，人の感情モダリティについて興味深いのは，印象を統合するだけでなく，感覚そのものが共起されることがある点である．有名な事例として，共感覚が挙げられる．一部の人は，文字を見ると対応する色が感じられることが報告されている[23]．これを色字共感覚という．既知の文字にはそれぞれ独立した色がある場合や，既知の文字は全て同じ色として認識される場合など，その性質には個人差がある．人によっては，数字には色があり，その色を元に演算することができる人もいる．

3 人工知能のモダリティ

人工知能とモダリティ

人工知能は 1957 年にダートマス会議で名称が決定する前後から，3 度のブームを起こしてきた．

1 回目のブームは人工知能が誕生した 1950 年代に起こり，サイバネティクスと合わせて，推論や探索など，様々な活用が期待された．ところが，難しい問題を解くことができないと指摘され，ブームは下火となった．それ以降，人工知能の分野では，人工知能そのものを研究するのではなく，人の知的な処理の一部を再現することが研究の主目的となり，問題を細分化して解決することに注力してきた．例えば，画像認識では画像に写っている対象を認識する，音声認識では音声に含まれる言語情報をテキストに変換して処理する，といった具合である．こうした背景から，これまでの人工知能はユニモーダルな研究が主流だった．

こうして各分野に特化した研究が続き，2012 年の深層学習を皮切りに，3 回目の人工知能ブームが起きた．深層学習では大規模なモデル，大量のデータを強力な計算機パワーで学習することで，高性能な画像やテキストの認識や生成を可能にし，限られた環境でならば人と遜色ない識別能力を有するようになった．

特に，ニューラルネットワークを使った機械学習では，入力情報を潜在的な特徴空間で表現し，分類を出力できる．この処理は人の感覚器からの入力を知覚し，クオリアを形成し，言語化する過程に酷似している．また，画像認識に使う CNN（Convolutional Neural Networks）は，入力画像に含まれる特定の形状に反応する視覚マスキングのようなニューロンを形成したり，おばあちゃん細胞仮説のように猫にだけ反応するニューロンを形成したりする現象[24]が確認されている．

こうした背景から，1 回目の人工知能ブームで期待されたように，人と同様にマルチモーダルな処理によるより高性能な識別や生成を目指すのは自然な流れだと思われる．2 章で述べたように，人の認知は複数の感覚モダリティで同一の情報が来るとその認知を高めるように働く性質がある．深層学習にもこれを組み合わせ，より高性能な処理を目指す研究が提案されてきている．

マルチモーダル深層学習の研究事例

機械学習分野において，マルチモーダルの発想自体は昔から存在していたが，深層学習が登場し，ここ 10 年で急速に発展してきた[25]．そのため，現在研究されている手法は比較的新しい技術やアプローチが多く，一概した定義というものは難しい．共通するのは，画像やテキストなど，複数のモダリティを扱うということである．その処理の一例を図 3 に示す．

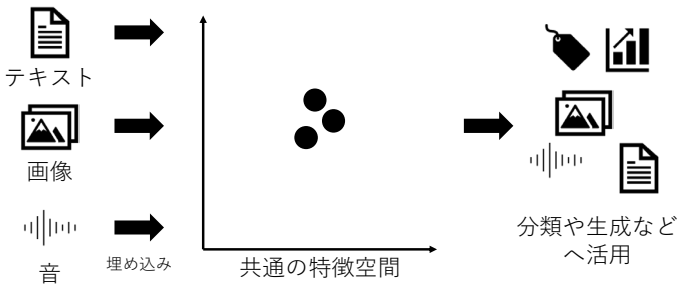

図 3 マルチモーダル深層学習の処理の一例

マルチモーダル深層学習では，テキストや画像など，複数のモダリティを同一の特徴空間に統合して利用する．これにより，テキストから得られた情報を画像へ変換したり，画像から得られた情報を音に変換したりと，異なるモダリティへの変換が可能となる．これらの処理において，特徴空間に上げることを埋め込み（embedding）やマッピング（mapping）と呼び，この特徴空間をマルチモーダル空間（multimodal space）や共通埋め込み空間（common embedding space），共通潜在特徴空間（common latent feature space）と呼ぶことがある．いずれにせよ英語での用語も対応する日本語訳もまだ確定していないものが多いが，マルチモーダルな情報を同一の特徴空間上で表現して扱っている点は共通している．

OpenAI 社が提案した CLIP（Contrastive Language–Image Pretraining）[26]では，画像とテキストの関連性を学習することで，画像からテキスト，テキストから画像への変換を可能とした．CLIP では，あらかじめ画像とそれに対応するテキストのデータセットを学習しておき，画像とテキストのペアの類似度を学習する．この情報を利用することで，新しいデータに対して画像とテキストによる補完を行うことでゼロショット学習を可能にしており，新しい画像分類タスクにおいても追加の学習なしで高い精度を発揮する．

また，大規模言語モデル（LLM: Large Language Model）の生成能力を補助する手法である RAG（Retrieval-augmented generation）にマルチモーダルな入力を用いる研究[27]もある．以前より LLM では，入力に完璧に答えようとするあまり，本来実在しない情報をあたかも本当にあるかのように提示してしまうため，その生成したテキストの信憑性が問題視されていた．そこで，外部のデータを使ってその信憑性を上げる RAG という手法が提案されていた．これにマルチモーダル情報を用いた研究[27]では，LLM の生成するテキストとは別に，外部からテキストや画像，音声などのマルチモーダルな追加データを入力とすることで，その信頼性を上げている．

これらの試みにより，今後は入力，出力ともにマルチモーダルな深層学習が主流となることが予測される．また，人で言う共感覚のようにクロスモーダルな処理を行うことで新しい発見が生まれるかもしれない．

4 まとめ

本稿では人の感覚のマルチモーダルと機械学習におけるマルチモーダルについて取り上げて紹介をした．

人の感覚モダリティは，複雑に絡み合っており，互いに影響することでその認知の補強を行なっていることがわかった．味と匂いが一致すると味が強化され，単語の意味と特徴が一致すると識別速度が上がるなどの事例が挙げられる．対して，認知が一致しないとき，McGurk 効果のように誤認をしたり，不気味の谷現象のように違和感を強く覚えたりすることがある．

人工知能，特に人の神経細胞の働きを模倣したニューラルネットワークでは，クオリアやおばあちゃん細胞のように，人の脳と興味深い共通点を見せた．そして，人と同様にマルチモーダルな情報をうまく扱うことで，分類精度の向上や生成のクオリティを上げることに成功している．

今後も，人の認知と人工知能の処理，２つのシステムを見比べることで，新たな知見や技術が生まれる可能性が考えられる．

参考文献

1. 北原義典，東倉洋一．音声の韻律情報と感情表現．電子情報通信学会技術研究報告; 1989; SP88-158, pp.27-32.
2. 重野純．感情を表現した音声の認知と音響的性質．心理学研究．2004; 74 (6), pp.540-546.
3. Scherer, K. R., & Zentner, M. R. Emotional effects of music: Production rules. In P. N. Juslin & J. A. Sloboda (Eds.). Music and emotion: Theory and research. Oxford University Press. 2001; 361–392.
4. E. Bigand, B. Poulin-Charronnat. Are we “experienced listeners”? A review of the musical capacities that do not depend on formal musical training. Cognition, 2006; 100(1): 100-130. https://doi.org/10.1016/j.cognition.2005.11.007.
5. Ronald E. Milliman. Using Background Music to Affect the Behavior of Supermarket Shoppers. Sage Publications, Inc. 1982; 46(3): 86-91.
6. Posner, M. I., Nissen, M. J., & Klein, R. M. Visual dominance: An information-processing account of its origins and significance. Psychological Review. 1976; 83(2), 157–171. https://doi.org/10.1037/0033-295X.83.2.157
7. Mehrabian, Albert. Silent Messages: Implicit Communication of Emotions and Attitudes (2nd ed.). Belmont, CA: Wadsworth. 1981.
8. Gross CG. Genealogy of the "Grandmother Cell. Neuroscientist. 2002; 8 (5): 512–518.
9. Ou, L.-C., et al. (2004). A study of colour emotion and colour preference. Part I: Colour emotions for single colours. Color Research & Application, 29(3), 232-240.
10. Kaya, N., & Epps, H. H. Relationship between color and emotion: A study of college students. College Student Journal. 2004; 38(3), 396-405.
11. Roland W. Fleming, Christiane Wiebel, Karl Gegenfurtner. Perceptual qualities and material classes. Journal of Vision July 2013, 13(9).
12. Elliot, A. J., & Maier, M. A. Color Psychology: Effects of Perceiving Color on Psychological Functioning in Humans. Annual Review of Psychology. 2014; 65, 95-120.
13. McGurk, H., & MacDonald, J. Hearing lips and seeing voices. Nature. 1976; 264(5588), 746-748.
14. Campbell, R. The processing of audio-visual speech: Empirical and neural bases. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences. 1994; 343(1306), 71-78.
15. Stroop, J. R. Studies of interference in serial verbal reactions. Journal of Experimental Psychology. 1935; 18(6), 643-662.
16. MacLeod, C. M. Half a century of research on the Stroop effect: An integrative review. Psychological Bulletin. 1991; 109(2), 163-203.
17. Köhler, Wolfgang. Gestalt Psychology. New York: Liveright. 1929.
18. Occelli, Valeria; Esposito, Gianluca; Venuti, Paola; Arduino, Giuseppe Maurizio; Zampini, Massimiliano. The Takete—Maluma Phenomenon in Autism Spectrum Disorders. Perception. 2013; 42 (2): 233–241. doi:10.1068/p7357. ISSN 0301-0066. PMID 23700961.
19. Schifferstein, H. N., & Verlegh, P. W. The role of congruency and pleasantness in odor-induced taste enhancement. Acta Psychologica. 1996; 94(1-3), 87-105.
20. T. Narumi, S. Nishizaka, T. Kajinami, T. Tanikawa and M. Hirose. MetaCookie+. IEEE Virtual Reality Conference, Singapore. 2011; pp. 265-266, doi: 10.1109/VR.2011.5759500.
21. Masahiro Mori, et al. The Uncanny Valley [From the Field]. IEEE Robotics & Automation Magazine. 2012; 19(2). 98-199.
22. 伴祐樹．クロスモーダルインタラクション最前線．電子情報通信学会誌. 2021; 104(12).1271-1278
23. 浅野倫子．色字共感覚：色と文字と学習の結びつき．年度日本基礎心理学会第 2 回フォーラム　共感覚と色情報処理. 2019.
24. Jeff Dean, Andrew Ng. Using large-scale brain simulations for machine learning and A.I. 26 June 2012; [cited 12 Nov. 2024]. Available: https://blog.google/technology/ai/using-large-scale-brain-simulations-for/
25. Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui & Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets Survey. The Visual Computer. 2021; 38, 2939–2970.
26. Alec Radford, Jong Wook Kim, et al.: Learning Transferable Visual Models From Natural Language Supervision. . 26 Feb 2021; [cited 12 Nov. 2024] Available: https://arxiv.org/abs/2103.00020
27. Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, William W. Cohen: MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022; pages 5558–5570.