人工知能の物体識別名称への人間の順応力を活用した画像・音声インタラクションシステムの研究
A study on a visual-voice interaction system utilizing human adaptability to object identification names of artificial intelligence

山次 善太 †，服部 一宏 †，金野 武司 †
Zenta Yamazi,Kazuhiro Hattori,Takesi Konno
† 金沢工業大学 工学部 電気電子工学科
Kanazawa Institute of Technology, College of Engineering,
† Electrical and Electronic Engineering
c6300928@st.kanazawa-it.ac.jp,konno-tks@neptune.kanazawa-it.ac.jp

概要

本研究では人工知能が未学習の対象を正しく判断できるようにすることを人工知能にだけ任せるのではなく，人間の順応力を利用して対象を特定することのできるシステムを構築し，その効果を実験により検証することを目指した．この中で，人間と機械が互いの認識を共有するために必要となる機能とそのメカニズムを明らかにすることを目的とした．実験環境として，我々はたくさんのオブジェクトが写る画像から特定の対象を探し出す課題を開発し，実験を行なった．実験には 3 つの条件：画像を何ら操作しない条件，マウスを使って画像を拡大・縮小できる条件，そしてマウス操作と共に人工知能とのインタラクションも行なえる条件を用意した．結果，画像操作をしない条件に比べて，マウス操作をする条件では有意に対象を早く発見できることが確認できたが，人工知能とインタラクションした条件では，我々の期待に反して対象の発見にかかる時間は長くなる結果となった．この結果から我々は，音声インタラクションを通じた対象の特定において人間の順応力を人工知能が活用するためには，対象への名付けのプロセスを人工知能がたどることができるようになる必要があるという示唆を得た．

キーワード：人工知能, ヒューマン・エージェントインタラクション（HAI）, 画像認識, 音声認識, 語用論

1. はじめに

深層学習の発展によって，音声や画像の認識技術は飛躍的に向上している．スマートフォンの顔認証機能は既に実用化され，自動車の自動運転技術も日々実用化に向けた研究・開発が進められている．しかし，自動運転車が実社会を走行する際にはまだ多くの困難があるようである．米国では，2021 年 7 月から 2022 年 5 月までの間に，自動運転レベル 2 で 367 件，自動運転レベル 3,4 で 130 件の事故が起きたことが報告されている [1]．具体的には，例えば 2016 年 5 月に起きた死亡事故では，大型トレーラーの側面に太陽光が反射した結果，状況の認識に誤りが生じたようである．また，2018 年 3 月に起きた試験走行中の自動運転車が歩行者をはねて死亡させた事故では，その原因として自転車を押した歩行者を歩行者として認識できなかったことが事故の原因と推定されている．

このように，カメラ画像から取得される情報から状況を認識する技術にはまだ多くの問題があり，様々な解決が試みられている．しかし，周囲 360 度の映像やレーダーなど，人間が知覚できない情報さえも人工知能が取得できるようになったとしても，実用化への難しさの本質的な問題は，人間社会の中では例外的な状況がいくらでも生じることにあるのではないだろうか．そういった例外的な状況に対しては，人間とのインタラクションの中で，人工知能が人間にとっての状況の意味を理解できるようになる必要があると考えられる．これに対処する方法として第一に思い浮かべられるのは，状況に応じて人工知能が学習をやり直すことであるが，我々はその過程が必ずしも必要とされるものではないのではないかと考えた．

例えば，人工知能がある情景から対象を抜き出し，人間が注目する対象を把握する場面を想像してみる．その対象が事前に学習された対象ではなかった場合，深層学習器は，人間の認識とは別の名前でその対象を呼ぶかもしれない．そのような事態への対応策として，人間と同じ認識の名前を再学習させること [2] が考えられるが，語彙の少ない子どもが大人と対話する

場面では，対象の呼び名が違っただけで，直ちに同じ対象に注目できなくなったりはしない．そこでは，大人側が子どもの呼び名を使うことで，注意の焦点を合わせ続けることができる．人工知能とのインタラクションにおいても，そういった人間側の対応の変化を前提にすることで，人工知能に学習能力を持たせずに対処できるケースがあるかもしれない．このような考えから本研究では，人工知能にのみ状況認識を任せるのではなく，人間との簡単なやりとりによって人間を適切にサポートするシステムを構築することができるのではないかと考えた．この中で，人間と機械が互いの状況認識を共有するために必要となる機能とそのメカニズムを検討・検証することを目的とした．

2. 実験課題の開発

前節で述べたようは考えを検証するための実験環境として，我々は画像を介して人間と人工知能がインタラクションする環境を考えた．これは，自動運転での対象認識や，多数の MRI 画像から病巣を発見するような医療診断場面，あるいは多くのモニターに表示された経済指標の把握といった状況に対応すると考えられる．具体的には，たくさんのオブジェクトが配置された画像を用意し，その中から特定の対象を探し出す課題[1]を考案した．この環境において，特定のオブジェクトを探し出すまでに掛かった時間や個数を計測することで，その課題のパフォーマンスを定量化した．また，パフォーマンスの検証を行うために，実験では以下の 3 つの条件を設定した．

人工知能とのインタラクションなし

- 条件 1：画像操作なしに提示された画像から特定のオブジェクトを探し出す
- 条件 2：画像の拡大，縮小および平行移動をマウス操作で行なう

人工知能とのインタラクションあり

- 条件 3：条件 2 に加えて，人間の音声発話を認識し，その結果を画面に表示する

2.1 人工知能の設計

人間とインタラクションする人工知能は，人間からの情報の受け取りには音声認識を用い，逆に人工知能から人間への認識の伝達には画像中に表示される認識エリアを示す四角形の描画と，認識単語の表示を行なうようにした．また条件 3 の実験参加者には，「○○を探して」と言えばそれを探してくれることを説明するようにした．

このようなインタラクションを実現するため，まず逐次的な画像認識器として Google の TensorFlow による SSD MobileNet v2 320×320 モデルを用いた．このモデルには 91 個のオブジェクトが学習されているが，本研究では対象の誤った認識を人工知能が提示するようにするために，参加者が探し出すオブジェクトには未学習のものを設定した．人間の発話音声の認識には Google Speech Recognition を使用し，そこで得られたテキストには MeCab[3] による形態素解析を実施した．切り出された名詞単語が認識する画像の中にあれば，そのオブジェクトを四角で囲み，その近傍に認識単語を表示するようにした．ただし，これだけでは人間が発話した単語しか認識結果として表示されないため，加えて WordNet[4] を用いて類義語を検索し，画像認識の結果に該当するものがあればそれを提示するようにした．これにより，例えば参加者が「チャリ探して」と言った場合に，認識結果として「自転車」と表示されるようになっていた．さらに，類義語の検索によっても該当する対象がない場合には，その時点で認識されているオブジェクト全てを四角で囲み，その認識名を表示するようにした．これらの仕組みにより，我々は参加者に「この対象はこういう名前で認識しているのだな」という気づきを与えることができるのではないかと予想した．

2.2 システム開発と実験設備

画像を映し出すモニター（iiyama 製 ProLiteXUB2790HS）の画面サイズは 597.9 × 336.3[mm] であり，画像はこの画面いっぱいに広がるように表示した．モニターに表示する画像は拡大・縮小の操作を行なうため，ベクターフォーマットで用意した．これにより，画像の拡大で画質が粗くなるようなことはなかった．また，画像の提示や画面の操作，および画像認識と音声認識を統合的に行なうためのシステム構築には，python に PyQt を組み合わせたプログラミング環境を利用した．

2.3 評価実験

実験の参加者は全員が金沢工業大学の学生で男性 13 名，女性 2 名の合計 15 名だった（平均年齢 21.53, $SD = 0.64$）．実験は全て同大学の研究室で実施した．実験で提示する画像には 41 種類のオブジェクトを合

[1]「ウォーリーを探せ」に類似した状況であると説明すれば理解しやすいだろう．

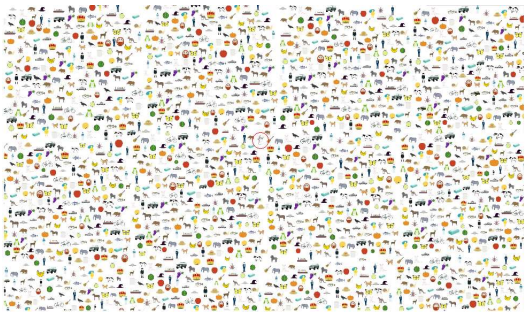

図 1　参加者に提示された画像例

計 1930 個使用し，ぞれぞれ配置の異なる画像を 5 枚用意した．一例として，1 つの画像を図 1 に示す．画像の初期ピクセルサイズは 1896×1030 だった．

参加者には，1 回の実験で 5 枚の画像が提示された．1 枚の画像に対して探す対象（同じもの）を 5 つ配置した．5 種類の探し出す対象は，それぞれ人型ロボット，犬，カブトムシ，カエル，ペットボトルだった．これらを参加者は制限時間 3 分以内に探し出すことが求められた．参加者が指定された対象を見つけた時には手元のマウスの左ボタンをクリックすると，画面には赤い丸印が現れるようになっており，直前の印は A キーを押すことで消せるようになっていた．参加者は操作を簡単に練習してから課題に取り組んだ．

3.　実験結果

画像の拡大縮小を行なわずに特定の対象を探した条件 1 での実験（操作なし条件），およびマウスで拡大縮小を行い対象を探した条件 2 での実験（マウス操作条件），そして人工知能とのインタラクションを行ない対象を探した条件 3 での実験（音声インタラクション）それぞれにおける，特定することのできた対象の数および，その特定までにかかった時間のグラフを図 2 に示す．

結果，探し出すことのできた対象の平均個数はそれぞれ 3.52, 4.20, 4.04 個であり，特定できるまでにかかった平均時間は 101.4, 76.7, 89.6 秒だった．それぞれについて一要因の分散分析を行なうと，個数については有意な主効果がなく（$F(2,72) = 1.85$, $p = .165$），時間については有意な主効果が確認された（$F(2,372) = 5.03$, $p = .007$）．多重比較の結果，条件 1,2 の間にのみ有意差があった（$p = .004$）．これらの結果は，参加者はマウス操作できる条件（条件 2）の方が，できない条件（条件 1）に比べて対象を短時間で探すことができたことを表すと共に，インタラクションできる条件（条件 3）は，条件 2 に比べてパフォーマンスが悪化したことを表している．

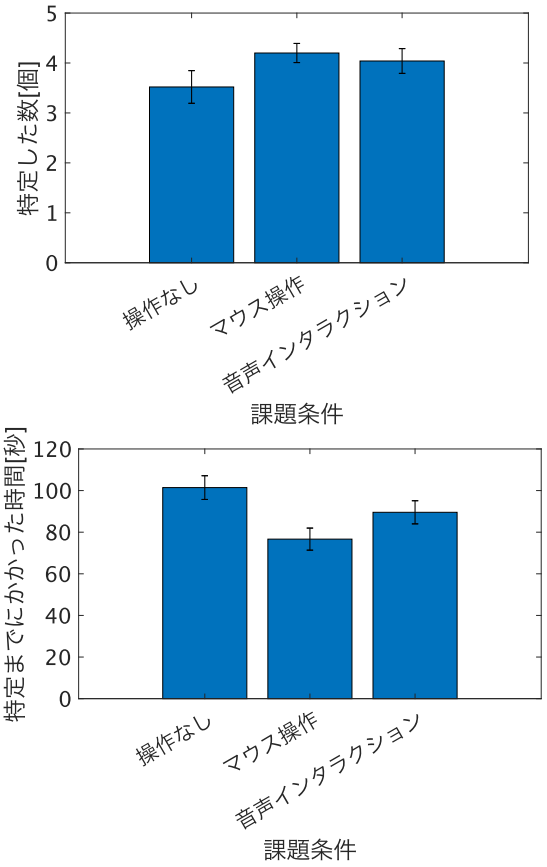

図 2　条件ごとでの特定した対象の数（上）と，その特定までにかかった時間 [秒]（下）

4.　議論

画像に対して何の操作も行なえない条件 1 よりも，マウスの操作によって画像を拡大・縮小したり，平行移動させることができる条件 2 の場合の方が，発見することのできる対象の数は多くなり，その特定にかかる時間は有意に短縮することが確認できた．この結果は，我々が開発した実験課題が，パフォーマンスの変化を定量的に計測できることを示している．

この実験課題において，人工知能とインタラクションした条件 3 のパフォーマンスは，我々の予想に反して悪化する結果となった．条件 3 におけるインタラクションの内容を個別に調べてみると，我々が想定するように，人工知能が通常とは異なる名前で表示（例えば人型ロボットを人間と表示するなど）したオブジェクトに対して，その呼び名を使って人工知能に話しかけた参加者が少なからずいた．しかし，我々が作成したシステムでは一定時間間隔で常にフレームに収められた画像を分析するようにしたため，画像の拡大・縮小操作によって対象の認識結果が変わることが度々あったようである．これによって参加者は，発するこ

とばをまた別のもの（例えば，種類や色の発話）に変更したり，あるいは発話を止めたりするなどしてしまっていた．このインタラクションにおいてまさに生じていたのは，1 つの対象を何と呼ぶかを共有する過程であり，システム側が用いたことばを人間が使用するところまでは起こっていたことがうかがえる．このときシステムには，自分が持っている対象と名前の関係性が人間に採用されたことを認識・保存する仕組みが必要なのだと思われる．当然のことながら，人間が種類や色といった，直接の名前から離れてしまった場合の対応を組み入れる必要はあると思われるが，直近の課題としては，一度認識された対象の名前を変更しない仕組みをシステムに導入したときに，課題のパフォーマンスが向上するのかを検証することが考えられる．

5. 結論

本研究では，不可避的に生じる人間社会での例外的な状況に人工知能が対応するための方略として，人工知能の学習に頼るのではなく，人間側が発揮する順応力を利用することを考えた．この考えの下で，人工知能に必要とされる機能やメカニズムを検討するために，我々は多くのオブジェクトが映る画像から特定の対象を見つけ出す課題を開発し，実験を実施した．結果，見つけ出すことのできた対象の数や，見つけ出すまでにかかった時間を測定することで，課題のパフォーマンスが定量的に計測できることを確認した．

この実験課題において，我々は人間の発話を認識し，そこから類推される対象を画像から探し出し，認識結果を画面に表示する仕組みを持った人工知能を開発した．そして，人工知能が持つ画像認識器が未学習の対象を探し出す課題を設定することで，人工知能にとっての例外的な状況を作り出した．この人工知能と共に課題に取り組んだ参加者は，例えば「人型ロボット」を探す中でそれを「人間」と認識するような人工知能を相手にして，いっときは「人間」を探せと指示することが確認された．しかし，我々が用意した人工知能は，対象の呼び名を自分に合わせてきたことを何ら認識せず，ただ画像の解析を行なって認識結果を変えてしまうことで，結果的に課題のパフォーマンスを悪化させることになったことがわかった．この結果は，人間の順応力の利用には，対象への名付けのプロセスを人工知能がたどれるようにする必要があることを示唆している．

謝辞

本研究は，JSPS 科研費基盤研究 (A)「道徳的行為者のロボット的構築による＜道徳の起源と未来＞に関する学際的探究」/課題番号 19H00524 の助成を受けた．ここに記し謝意を表します．

文献

[1] 米運輸省道路交通安全局（NHTSA）. Incident Reporting for Automated Driving Systems and Level 2 Advanced Driver Assistance Systems. Standing General Order 2021-01, https://www.nhtsa.gov/laws-regulations/standing-general-order-crash-reporting.
[2] 長井隆行, 青木達哉, 中村友昭 (2016). 言語を理解するロボット実現への確率ロボティクスアプローチ. システム制御情報学会誌,Vol.65,No.12,pp.32–38.
[3] 工藤拓, 山本薫, 松本祐治 (2004). Conditional Random Fields を用いた日本語形態素解析. 情報処理学会研究報告自然言語処理（NL）,pp.88–96.
[4] 福本文代, 鈴木良弥 (2002). WordNet の同義語クラスとその上位関係を利用した文書の自動分類. 情報処理学会論文誌,Vol.43,No.6,pp.1852–1865.