人の「示す」動作を考慮して AI モデルを構築するシステム
――AI モデルの開発をより手軽により正確に――

１．発表者：

矢谷　浩司（東京大学 大学院工学系研究科電気系工学専攻　准教授）

２．発表のポイント：

- ユーザが行う物体を示す動作（教示動作）を基に、撮影された静止画においてどの部分にユーザが AI モデルに学習させたいと思う物体があるかを推定し、その部分に重きをおいて学習を行うことができる対話型 AI モデル構築システム LookHere を構築しました。
- 持ち上げて見せる、指し示すなど、人間が自然と理解できるユーザの教示動作を、AI モデルの学習において明示的に考慮できるシステムを実現し、ユーザが意図する物体を正確に AI モデルの学習に取り入れることを可能としました。
- 本研究成果は AI の専門家や開発者でない一般のユーザが信頼性のある AI モデルを構築することを助け、AI 開発の市民化を加速させることが期待されます。

３．発表概要：

東京大学大学院工学系研究科電気系工学専攻の矢谷浩司准教授らのグループは、ユーザが行う教示動作を基に、撮影された静止画においてどの部分にユーザが AI モデルに学習させたいと思う物体があるかを推定し、その部分に重きをおいて学習を行うことができる対話型 AI モデル構築システム LookHere を構築しました。このシステムにより、ユーザはシステムが学習させたい物体部分が AI モデルの学習に使われるかを事前に確認することができ、信頼性のある AI モデルの構築に繋がります。さらに、LookHere システムを実現するために収集した、ユーザの教示動作の画像データセット HuTics も公開しており、ユーザインタフェースや人工知能の研究者・技術者が自由に使用して、新たな技術やアプリケーションを構築していくことが期待されます。

本研究成果は、2022 年 10 月 31 日（米国太平洋夏時間）に国際会議 The ACM Symposium on User Interface Software and Technology において口頭発表とデモを行いました。

４．発表内容：

人工知能技術（Artificial Intelligence、AI）は私達の生活のなかに急速に取り入れられています。一般のユーザも手軽に AI モデルを構築できる方法として、対話型 AI モデル構築（Interactive Machine Teaching）というものがあります。このような対話型 AI モデル構築の具体的な例としては、コンピューターに接続されているカメラの前で、認識させたい物体や動作をユーザ自身が提示するシステムがあります。システムはその様子を静止画や動画で撮影し、そのデータを AI の学習に利用します。ユーザは AI に認識させたい物体や行為を例示して「教える」だけで、通常なら必要とされる技術的知識や経験がなくとも、自分自身の AI モデルを作り出すことができます。

一方で、このようなシステムにおいては、AI モデルが構築される際に、提供した写真や動画のどの部分が学習に利用されたかが明らかではないため、ユーザの意図しない部分が学習に使われ、構築された AI モデルが十分に信頼できないものになってしまうことがあります。例え

ば、ユーザが非接触式の体温計を教示し、それを認識する AI モデルを構築しようとする場面を考えます（図 2a）。この様子を人間が見ると、このユーザが非接触式の体温計を教示したいことはすぐに分かります。しかし、持ち上げて見せる、指し示すなど、人間が自然と理解できるユーザの教示動作を、AI モデルの学習においては必ずしも明示的に考慮しないため、ユーザが認識させたい部分とは違う部分を学習に利用してしまうことがあります。図 2b はこのシーンにおいて AI モデルが認識に特に利用している部分を可視化したものですが、非接触式の体温計ではなく背景の部分がモデル内で特に利用されていることを示しています。そのため、この AI モデルは例えばユーザの背景が変わってしまうと、うまく認識ができないなどの問題が起こりえます。

そこで、本研究ではユーザが行う教示動作を基に、撮影された静止画においてどの部分にユーザが AI モデルに学習させたいと思う物体があるかを推定し、その部分に重きをおいて学習を行うことができる対話型 AI モデル構築システム LookHere を構築しました（図 1）。このシステムでは、ユーザが AI モデルに学習させたい物体をカメラの前で教示すると、どの部分が学習させたい物体であるとシステムが認識しているかをハイライトします（図 2c）。これによって、ユーザはシステムが学習させたい物体部分が AI モデルの学習に使われるかを事前に確認することができ、信頼性のある AI モデルの構築に繋がります。ユーザの教示動作を認識するために、教示動作を収集した画像データセット HuTics も構築しました。このデータセットに基づいた認識モデルにより、Exhibiting、Pointing、Presenting、Touching という 4 つの教示動作を LookHere では認識できるようになっています。

実験では、LookHere を使用する場合と使用しない場合、さらに使用しない場合においては、学習させたい物体が画像中のどこにあるかの情報付けを実験参加者が手動で後から行う場合とを比較し、AI モデルの学習に使われるデータ構築にかかる時間、及びそのデータを用いたモデルの学習対象の物体部分を画像中から切り出す（セグメンテーション）精度を検証しました。結果、LookHere を使用する場合は、手動で情報付けを行う（アノテーション）場合と比較して、作業時間を 11〜14 分の 1 程度に削減できました。また、LookHere を使用せずアノテーションがない場合、セグメンテーションの精度は 13.9%でしたが、LookHere の場合には 60.5%となり、LookHere を使用せずアノテーションを行った場合の 72%程度に匹敵する精度となりました。この実験により、LookHere によって、ユーザが手動で情報付けを後から行わずとも、信頼性の高い AI モデルを構築できる可能性を示しました。

また、LookHere システムを実現するに当たり構築した HuTics を一般に公開しています。このデータセットには Exhibiting、Pointing、Presenting、Touching という 4 つの教示動作を撮影した合計 2,040 枚の静止画が含まれています（図 3）。このような人間の教示動作を網羅的に含めたデータセットは世界初であり、このデータセットを使うことにより、例えばユーザの教示動作で示された物体にフォーカスを当てるような視覚効果を実現するカメラや、オンラインミーティングにおいてユーザが手に持っている物体にバーチャル背景が邪魔をしないようにすることなどが実現されることが期待されます。

５．発表詳細：

国際会議名：The ACM Symposium on User Interface Software and Technology（UIST）
論文タイトル：Gesture-aware Interactive Machine Teaching with In-situ Object Annotations
著者：Zhongyi Zhou*, Koji Yatani*

プロジェクトホームページ：https://zhongyi-zhou.github.io/GestureIMT/
デモンストレーション動画：https://www.youtube.com/watch?v=ZS-Ser7-vGI

６．問い合わせ先：

＜研究に関すること＞
東京大学 大学院工学系研究科 電気系工学専攻
准教授　矢谷 浩司（やたに　こうじ）

＜報道に関すること＞
東京大学 大学院工学系研究科 広報室

７．添付資料：

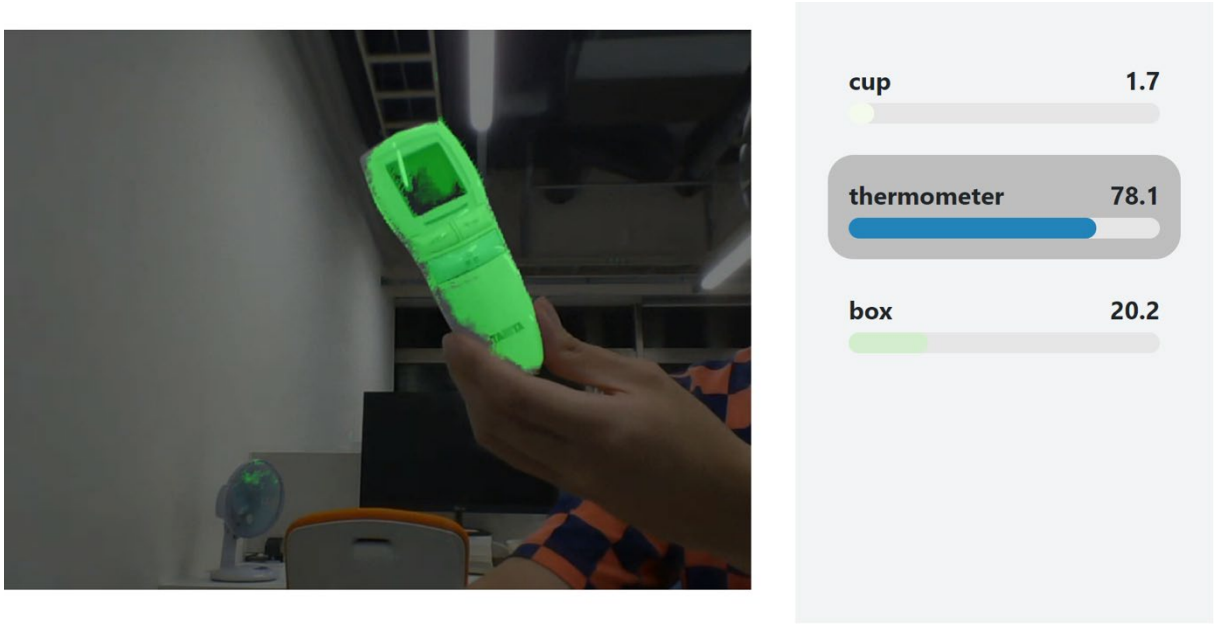


図 1. LookHere システムのインタフェース。ユーザが AI モデルに学習させたいと思う物体をカメラの前に提示すると、ユーザの教示動作を基に AI モデルの学習に使うべき物体の領域を推定し、この図のようにユーザに提示する。ユーザは緑色のハイライトが学習させたい物体の上にあることを確認し、撮影を行う。これにより、ユーザは画像中のどの部分が AI モデルの学習に特に使用されるかをデータ収集時に確認することができ、信頼性のある AI モデルの構築を追加の負荷（手作業でのアノテーション等）なく実現できる。

図 2a. 非接触式の体温計を認識させるためにカメラに提示しているところ。

図 2b. 図 2a と同じ画像において、LookHere を使用しない場合に AI モデルの学習において、特に使用された部分を緑色でハイライトしたもの。非接触式の体温計の部分以外にも、背景にある三脚や棚、ホワイトボードなど、無関係な部分も学習に使用されている。このため、例えば違う背景で同じ AI モデルを使用した場合、想定通りに認識しない可能性がある。

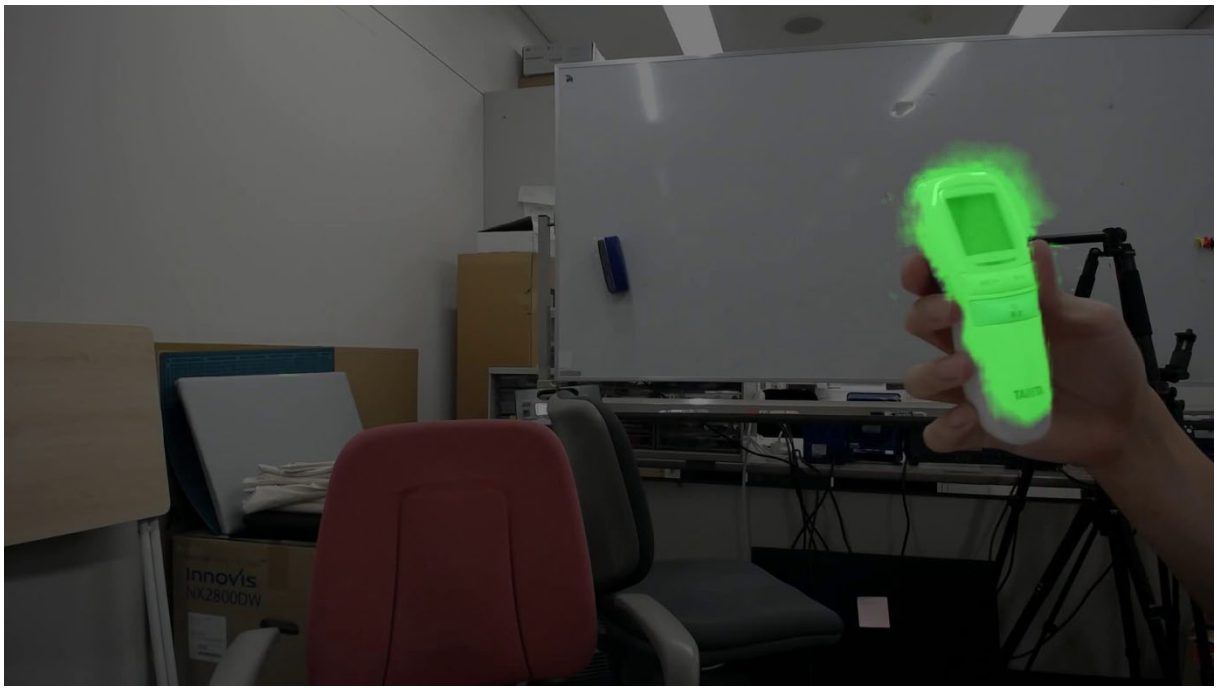
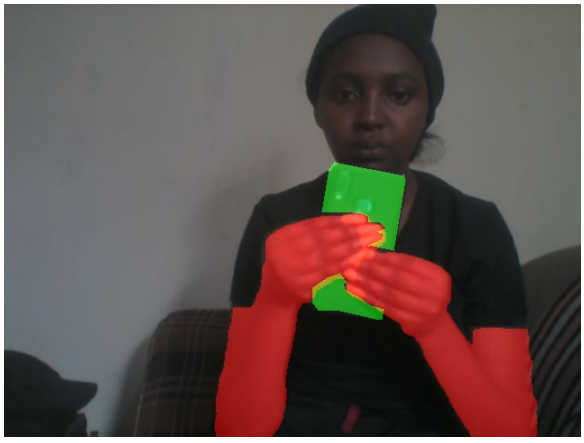


図 2c. 図 2a と同じ画像において、LookHere を使用した場合に AI モデルの学習において、特に使用された部分を緑色でハイライトしたもの。ユーザの教示動作（この場合、物体を保持する動作）から、非接触式の体温計の部分を正確に切り出し、この領域に重点を置いて学習が行われ、背景が変更されても AI モデルが想定通りに認識を行うことが期待できる。

Exhibiting

Pointing

Presenting



Touching

図 3. ユーザの教示動作を収集した画像データセット HuTics 内の画像一例。画像はそれぞれ Exhibiting、Pointing、Presenting、Touching という 4 つの教示動作を示す。緑色でハイライトされている部分はユーザが教示している物体の領域、赤色でハイライトされている部分はユーザの手と腕となる。HuTics には合計 2,040 枚の静止画が含まれており、このような人間の教示動作を網羅的に含めた世界初のデータセットとなる。