arXiv:2304.14501v1 [cs.CV] 7 Mar 2023

读懂我的心：用于人类信念预测的多模态数据集。

Jiafei Duan[1], Samson Yu[2], Nicholas Tan[2],Yi Ru Wang[1] 和 Cheston Tan[3]

摘要—理解人类意图是实现协作场景中高效、有效人机交互（HRI）的关键。为推动人工智能（AI）系统推断人类信念能力的发展与评估，我们基于物体-情境关系，推出了一个用于意图预测的大规模多模态视频数据集。

I. 引言

具备平均水平社会认知能力的人能够仅凭社会互动过程中呈现的非语言交流信号（例如，注视、手势、姿态和情境信息）推断他人的信念 [1], [2], [3], [4]。机器人学和人工智能领域近期的技术进步，提升了各领域（例如，制造业、服务业、医疗保健等）对人机交互（HRI）的需求 [5]。理解人类意图的关键在于感知互动中的人类动作，并理解物体-情境关系，即对特定情境中出现的物体的认知。因此，我们提出了一个新颖的视频数据集，其中包含多模态标注，能够捕捉物体-情境场景中的人类行为。

II. 数据收集

我们的数据集包含来自 15 种不同情境的 10 对参与者——其中五对是朋友，五对是陌生人。我们从 900 段第一人称视角和第三人称视角视频中采集了 347,490 个视频帧。如图 1B 所示，每对参与者中的两人分别获得情境物体（演示者）或情境专用工具（预测者）。这些物体和工具都是常见的家居用品，其中大多数都可以在 YCB 数据集中找到[6]。演示者会接到一项隐式情境任务，并根据任务选择一个情境物体。随后，演示者使用所选物体，以非语言方式向预测者传达该隐式情境任务；预测者则通过推断演示者的意图，从情境专用工具列表中选择一件工具。

在整个实验过程中，参与者都佩戴降噪耳机，并在信念更新时说出自己想到的物品名称。这使我们能够准确标注与所采集视频帧对应的参与者隐藏信念。我们还通过两个 Leap Motion 传感器收集了手势数据。其他输入模态，包括物体检测、姿态估计和注视追踪，均通过后处理方法实现。我们使用 Detecto [?] 检测物体，并在自标注的视频帧上进一步微调，使用 Gaze360 [7] 收集三维人类注视方向，并使用 OpenPose [8]获取姿态估计所需的所有关键点。

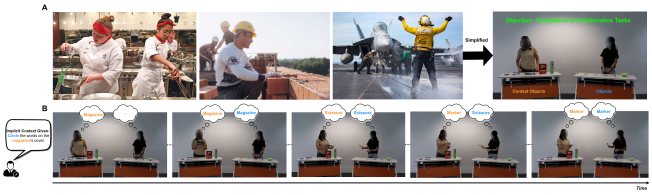


图 1：（A）需要通过非语言交流进行信念推断的现实世界协作任务示例。（B）数据收集期间提供的指示示例。

III. 数据集分析

该数据集包含 900 段录制视频，均从自我中心第三人称视角同步拍摄，共有 347,490 个标注帧，视频总时长约为 3.2 小时。如图 2B 所示，上下文物体序列的平均时长约为 8 秒，即在 30fps 帧率下约为 250 视频帧。所有视频均录制了音频，且音频已同步。此外，图 2A 展示了所有参与者配对中匹配的频率以及可能的物体-上下文配对。利用从训练集获得的数据，我们可以创建一个物体-上下文关系矩阵，精确描绘相互关联的上下文与物体关系。该矩阵还可在训练中用作先验知识，表征上下文与物体之间的功能或用途关系。

IV. 结论

我们引入了一个具有挑战性的多模态数据集，以支持在物体-上下文场景中根据视频进行人类信念预测。我们希望这项工作能够促进未来对人类信念的理解研究，并为人机交互领域作出贡献。

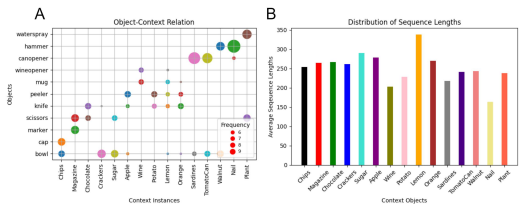


图 2：数据集统计概览。(A) 所有潜在物体-上下文匹配的频率。(B) 各上下文物体类别的序列长度分布。

[1]美国华盛顿大学，duanj1@cs.washington.edu[2]新加坡新加坡国立大学[3]新加坡科技研究局前沿人工智能研究中心

参考文献

[1] Robert A Hinde。《style id='2'>人类社会行为的生物学基础。McGraw-Hill，1974。 [2] Henry M Wellman、David Cross 和 Julanne Watson。心智理论发展的元分析：关于错误信念的真相。《style id='5'>儿童发展，72(3):655–684，2001。 [3] Rebecca Saxe。独特的人类社会认知。《style id='8'>神经生物学最新观点，16(2):235–239，2006。 [4] Yi Ru Wang、Jiafei Duan、Sidharth Talia 和 Hao Zhu。交互式人工智能与人类舒适度的研究。《style id='11'>arXiv预印本 arXiv:2302.14360，2023。[5] Jiafei Duan、Samson Yu、Hui Li Tan、Hongyuan Zhu 和 Cheston Tan。具身人工智能综述：从模拟器到研究任务。《style id='14'>《IEEE计算智能新兴主题汇刊》，6(2):230–244，2022。[6] Berk Calli、Arjun Singh、Aaron Walsman、Siddhartha Srinivasa、Pieter Abbeel 和 Aaron M Dollar。YCB物体与模型集：迈向操作研究的通用基准测试。载于《style id='17'>2015年国际先进机器人会议（ICAR），第510–517页。IEEE，2015。[7] Petr Kellnhofer、Adria Recasens、Simon Stent、Wojciech Matusik 和 Antonio Torralba。Gaze360：自然环境下不受物理约束的视线估计。载于*IEEE/CVF*国际计算机视觉会议论文集，第6912–6921页，2019。 [8] Zhe Cao、Tomas Simon、Shih-En Wei 和 Yaser Sheikh。利用部件亲和场进行实时多人二维姿态估计。载于《style id='23'>IEEE计算机视觉与模式识别会议论文集，第7291–7299页，2017。