psifx - 心理与社会互动特征提取包

Guillaume Rochette* GUILLAUME.ROCHETTE@UNIL.CH
UNIL, Switzerland

Mathieu Rochat MATHIEU.ROCHAT@UNIL.CH
UNIL, Switzerland

Matthew J. Vowels* MATTHEW.VOWELS@UNIL.CH
UNIL, Switzerland
The Sense, CHUV, Switzerland

Editor:

Abstract

ψfx 是一个即插即用多模态特征提取工具包，旨在促进并普及最先进的机器学习技术在人文科学研究中的应用。它由以下需求驱动：(a) 自动化和标准化通常需要昂贵、耗时且不一致的人力劳动的数据标注过程；(b) 开发和分发开源社区主导的心理学研究软件；以及 (c) 为非专业人士用户提供大规模访问和易于使用的途径。该框架包含了一系列用于任务的工具，例如说话人分离、音频中的字幕转录与翻译；视频中多个人体态、手部姿态及面部姿态估计和注视跟踪；以及由大型语言模型支持的互动文本特征提取。该软件包采用模块化和任务导向的方法设计，使社区能够轻松添加或更新新工具。这种组合为心理学和社会科学研究中的实时行为现象深入研究创造了新的机遇。

Keywords: psifx, 多模态, 视频, 音频, 语言特征提取, python

1 介绍

人类和社会互动的研究需要访问有意义且可解释的特征来表示这些互动。目前，标准方法涉及人类观察编码，这带来了三个主要挑战：在时间和培训方面成本过高 (Bulling et al., 2023)，编码者之间缺乏标准化 (Harris and Lahey, 1982)，以及难以扩展到大型数据集。例如，仅手动转录就比音频本身的时长多出五到十倍的时间 (Bazillon et al., 2008)。这些问题阻碍了行为研究的进展，特别是在心理健康研究和干预设计方面。

*同等贡献。

为了得到一组表示复杂视频和音频数据的变量，可以将行为互动分解为可解释的特征。这些特征包括非言语行为（身体和头部姿势、动作、凝视、面部表情）、副语言特征（音高、语调）以及口头的特征（转录语言）。这些特征代表可以直接观察到的事件，与压力或情绪等更难以捉摸的心理结构形成对比。

尽管最先进的机器学习技术已公开用于这些任务，包括骨骼姿态估计 (Cao et al., 2018)、面部分析 (Baltrusaitis et al., 2018)、语音特征提取 (Eyben et al., 2010) 和自动转录 (Radford et al., 2023)，但在心理学和社会科学领域仍存在显著的采用障碍。这些障碍包括设置和部署的技术难题、对编程专业知识的需求以及使用第三方服务时关于数据隐私的担忧。

为了解决这些挑战，我们提出了 psifx（P 心理学和 S 社会 I 互动 F 特征 eX 提取），一个开源项目，提供：

- 一种综合方法以实现客观非言语、副语言和言语特征的自动化提取
- 高效并行和硬件加速处理大规模数据集
- 通过统一的命令行界面简化设置和使用
- 公共的、面向社区的存储库，包含免费使用的 Python 包和 Docker 镜像
- 使用人类可读的数据格式标准化任务输出
- 本地数据处理能力以保护敏感信息

2 功能与特性

psifx 实现了一种围绕三种主要模态组织的模块化架构：视频、音频和文本。每种模态都包含可以独立使用或组合成处理管道的专业模块。该系统设计为可扩展，允许添加新模块而不影响现有功能。

2.1 安装和使用

psifx 的一个重要优势在于其简化的安装过程。尽管许多开源项目的性能令人印象深刻，但仅安装一个这样的库所需的时长和专业知识可能相当可观。当希望同时安装多个库时，相关难度会增加，并且会出现包依赖之间的兼容性问题。确实，为一个（单个）知名的计算机视觉软件包进行安装程序可能会需要超过 20 条命令来安装一系列低级依赖项。此外，psifx 提供了一个可以通过 pip 安装的 Python 包，以及一组准备好的容器化镜像，在这些镜像中，外部库之间相互兼容。这提供了一种开箱即用的实用性，而许多开源项目则不具备这一点。

我们提供了一个简单的命令行界面来与该包进行交互。例如，要使用 mediapipe 提取姿态，可以使用以下命令：

```
psifx video pose mediapipe multi-inference \
    --video input.mp4 \
    --poses output/poses.tar.gz \
    -- masks MaskD
```

也提供了 Python 接口，尽管主要使用场景优先考虑没有编程经验的用户。

2.2 视频处理

视频处理流水线集成了多种先进的非语言特征提取工具：

2.2.1 多目标追踪

我们集成了 Samurai(Yang et al., 2024)，这是一个基于 Meta 的 Segment Anything Model 2(Ravi et al., 2024) 构建的对象跟踪算法，可以默认用于在视频中跟踪多个人，或用于跟踪特定对象类别。它与集成在 Ultralytics 包中的 YOLO(Jocher et al., 2023) 结合使用，以实现自动的人/物体检测和跟踪。

2.2.2 人体姿态估计

我们将 MediaPipe (MediaPipe, 2024) 结合起来，用于实时估计身体、面部和手部的配置。这使得能够追踪与心理治疗和行为分析相关的肢体运动和“体语”。其设计目的是可以利用跟踪算法中的掩码来进行多人姿态估计。

2.2.3 面部分析

OpenFace2.0 (Baltrusaitis et al., 2018) 集成提供：

- 目光估计
- 面部关键点
- 表情动作编码系统（FACS）单元
- 头部姿态估计

它被设计成能够利用跟踪算法中的掩码来进行多人姿态估计。

2.3 音频处理

音频管线提供了以下语音分析功能：

2.3.1 说话人分割与重新识别

我们集成了 pyannote (Bredin, 2023) 用于说话人分段，并使用集成嵌入模型实现定制的说话人重识别系统。这使得在多麦克风设置中能够可靠地将说话人映射到其原始音频通道。

2.3.2 转录与分析

该系统包含：

- 耳语 (Radford et al., 2023)，特别是耳语 (Bain et al., 2023) 实现，用于多语言转录
- 开放微笑 (Eyben et al., 2010) 用于副语言特征提取
- 增强转录结合说话人分离和说话人识别

2.4 文本处理

文本处理功能利用 LangChain (Chase, 2022) 提供：

- 灵活的 LLM 集成（本地或基于云）
- 基于通用指令的处理
- 交互式聊天功能
- 多个模型后端（Hugging Face, Ollama, OpenAI, Anthropic）

3 测试

该软件包包括用于自动集成测试的 CI/CD 工作流，以及 PyPI 和 Docker 发布。

4 数据质量和硬件指南

为了获得最佳性能，我们建议使用同步多摄像机/麦克风设置，并以单人视频帧进行姿态估计。麦克风应佩戴在身上，以便清晰区分说话者的声音。建议使用足够的摄像头分辨率和帧率，并控制或漫射照明以及音频条件。

该软件包支持仅使用 CPU 和 GPU 加速操作。推荐使用启用了 CUDA 的硬件进行大规模数据处理以及本地 LLM 托管，用于文本分析工具。

5 开发路线图

当前开发重点包括 GUI 设计与实现，生理传感的集成（例如从红外摄像机估算呼吸率），以及多模态说话人分离方法，以便在没有多个领夹麦克风的情况下也能成功进行说话人分离。

6 结论

我们介绍了开源项目 psifx，这是一个用于多模态估计与心理和社会科学相关特征的综合软件包。psifx 的目标是标准化和简化人类互动的注释过程，以提高该研究领域的可靠性和可重复性。同时通过消除设置以及使用便利性的多个技术障碍来简化并推广社区内最先进的机器学习技术的应用，同时保持效率。此外，开源和社区驱动的特点将有助于塑造和支持有机增长，并增加项目的长期发展能力。我们希望 psifx 能够为经验研究人员提供可用的、现代的、开放的和由社区驱动的非言语、副语言和言语特征提取工具。

参考文献

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023.

T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L-P. Morency. OpenFace 2.0: Facial behavior analysis toolkit. *13th IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

T. Bazillon, Y. Estè, and D. Luzzati. Manual vs assisted transcription of prepared and spontaneous speech. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008. doi: http://www.lrec-conf.org/proceedings/lrec2008/pdf/277_paper.pdf.

H. Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle benchmark, and recipe. *Proc. INTERSPEECH 2023*, 2023.

L. Bulling, R.E. Heyman, and G. Bodenmann. Bringing behavioral observation of couples into the 21st century. *Journal of Family Psychology*, 37(1):1–9, 2023. doi: 10.1037/fam0001036.

Z. Cao, G. Hidalgo, Simon T., S.-E. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *arXiv:1812.08008v1*, 2018.

H. Chase. Langchain, oct 2022. URL `https://github.com/langchain-ai/langchain`.

F. Eyben, M. Wollmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. *Proceedings 18th ACM international conference on multimedia*, (1459-1642), 2010. doi: 10.1145/1873951.1874246.

F.C. Harris and B.B. Lahey. Recording system bias in direct observational methodology: A review and critical analysis of factors causing inaccurate coding behavior. *Clinical Psychology Review*, 2(4):539–556, 1982. doi: https://doi.org/10.1016/0272-7358(82)90029-0.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. URL `https://github.com/ultralytics/ultralytics`.

MediaPipe, 2024. URL `https://developers.google.com/mediapipe/solutions/vision/pose_landmarker/`.

A. Radford, J.W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning, PMLR*, 202:28492–28518, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint*, arXiv:2408.00714, 2024.

Cheng-Yen Yang, Hsiang-Wei Huang, Wenhao Chai, Zhongyu Jiang, and Jenq-Neng Hwang. Samurai: Adapting segment anything model for zero-shot visual tracking with motion-aware memory. *arXiv preprint*, arxiv:2411.11922, 2024.