

# Artifacts Guide: Evaluation of Video-Assisted Annotation of Human IMU Data

1<sup>st</sup> Alexander Hoelzemann  
*Ubiquitous Computing*  
*University of Siegen*  
Siegen, Germany  
alexander.hoelzemann@uni-siegen.de

1<sup>st</sup> Marius Bock  
*Ubiquitous Computing & Computer Vision*  
*University of Siegen*  
Siegen, Germany  
marius.bock@uni-siegen.de

2<sup>nd</sup> Kristof Van Laerhoven  
*Ubiquitous Computing*  
*University of Siegen*  
Siegen, Germany  
kvl@eti.uni-siegen.de

**Abstract**—This artifact guide outlines the supplementary material for our paper *Evaluation of Video-Assisted Annotation of Human IMU Data Across Expertise, Datasets, and Tools*. The supplementary material comprises Python code that replicates our findings, along with instructions for downloading and deploying this code, and a list of hardware requirements for successful execution. Additionally, we present detailed results of our study including per-participant NASA-TLX questionnaire results and per-participant evaluation metrics. Lastly, we provide in this guide all necessary material and offer supplementary insights into the annotation tools employed in our study so that the study could be replicated by other researchers with additional participants. The code, data and supplementary information can be publicly downloaded from the following repository: [https://github.com/mariusbock/video\\_assisted\\_annotation](https://github.com/mariusbock/video_assisted_annotation)

## I. DOWNLOAD AND DEPLOYMENT

The relevant code and collected study data for our Paper *Evaluation of Video-Assisted Annotation of Human IMU Data Across Expertise, Datasets, and Tools* [1] is available in one location under a GitHub project repository<sup>1</sup>. The structure of this repository, once downloaded, is as follows:

- 1) *annotations*: Folder containing the exported annotations of each participant’s annotation sessions using either the MaD-GUI [2] or ELAN-Player [3] for annotation.
- 2) *sensor\_data*: Folder containing the raw inertial sensor data of the WEAR [4] and Wetlab [5] dataset used for each of the four types of sessions. The sensor data is by sessions, i.e., each file corresponds to one of the four ten minute segment of the WEAR or Wetlab dataset, which are to be labeled by participants.
- 3) *data\_prep.py*: Python script used for converting the extracted annotations to CSV format and combining them with the raw ground truth data of each dataset.
- 4) *main.py*: Main Python script to recreate all plots and reproduce evaluation results mentioned in the paper.
- 5) *requirements.txt*: Requirements file which serve as a list of items to be installed by pip, when using pip install.
- 6) *study\_results.xlsx*: Excel sheet containing all per-participant NASA and evaluation metrics results.
- 7) *nasa\_feedback\_template.pdf*: Post-experiment questionnaire sheet collecting NASA-TLX [6] results and feedback regarding the annotation tools.

<sup>1</sup>[https://github.com/mariusbock/video\\_assisted\\_annotation](https://github.com/mariusbock/video_assisted_annotation)

In order to reproduce our presented results please follow the step-by-step guide as depicted in the *README.md* file of the repository. Given that the main script is only creating plots and calculates evaluation metrics such as F1-score, our repository is able to be run on most personal computers and laptops. It is further not required to use a computer that contains a GPU.

**Expected Results:** Running the main script results in the creation of a *plots* folder and two JSON-formatted text files. The *plots* folder will contain two types PNG files, being color-coded visualization of each participant’s provided annotations (top plot) compared to the ground truth annotations (bottom plot) and confusion matrices calculated comparing participant and ground truth annotations. The color-coded visualization plots were cropped and combined during the creation of Figure 2 of the main paper. Each PNG file’s filename encodes all information which session it represents. The JSON-formatted text files contain per-session F1-score, Cohens- $\kappa$  and NULL-class accuracy. In order to check the validity of obtained results, we provide an excel sheet, *study\_results.xlsx*, containing all per-participant NASA and evaluation metrics results one obtains when running our repository.

## II. DATASET

The datasets utilized in our investigation include 20-minute synchronized inertial measurement unit and video recordings of a solitary participant for each dataset, WEAR and Wetlab. As a result, the dataset comprises data from 15 participants who annotated 40 minutes of data, resulting in 600 minutes of re-annotated activity data. Technical specifications for each dataset can be found in the original publications [4], [5]. Classes that were re-annotated during our studies are:

- 1) *WEAR* [4]: running (sidesteps), bench-dips, stretching (shoulders), jogging (butt-kicks), burpees, and lunges.
- 2) *Wetlab* [5]: pouring, pipetting, transfer, stirring, cutting, pestling, and mixing.

**Annotation Tools:** Participants were instructed to retrospectively annotate the data using two distinct software tools for this purpose: MaD-GUI and ELAN-Player, see Figure 1.

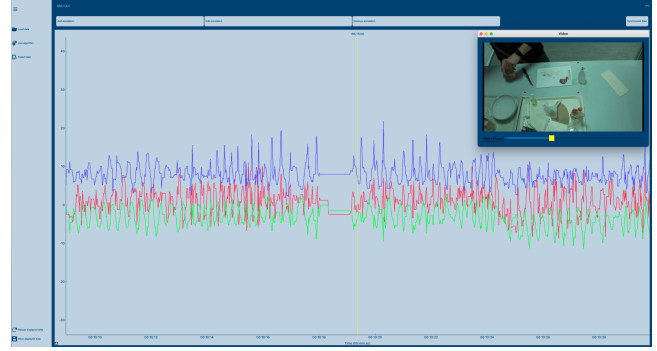
Developed by Ollenschläger *et al.* at Friedrich-Alexander-Universität Erlangen-Nürnberg in Germany, MaD-GUI is a tool designed to aid researchers in visualizing, synchronizing, and annotating inertial data. The software is available for

free access on GitHub<sup>2</sup> or can be directly installed using the Python package manager pip, following the instructions outlined in the tool’s documentation<sup>3</sup>. Its core strengths lie in the visualization of sensor data and annotations, both displayed prominently in the main frame. Users can add annotations by selecting from a predefined list. Additionally, the tool empowers users with the flexibility to set multi-tier labels.

ELAN is a video annotation tool that was initially developed for linguistic analysis at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands. It has since evolved to be used for a variety of annotation tasks, including annotating IMU data. ELAN enables researchers to synchronize IMU data with video recordings and annotate the data with timestamps and labels. The software can be downloaded from the ELAN-Player’s official website<sup>4</sup>. The primary interface of the tool presents the video footage, with sensor data occupying a supplementary role within the frame. Annotations are created manually within ELAN. Notably, similar to MAD-GUI, the tool allows for the construction of multi-tier label structures. To seamlessly import time-series data and video files into the MaD-GUI and ELAN-Player software, researchers should follow the provided guidelines by the original authors made available on YouTube<sup>5,6</sup> or as part of their online documentation<sup>7</sup>. These guidelines offer detailed instructions on the data format requirements, ensuring that the data meets the specific specifications of each tool. For MaD-GUI, the required format for the time-series data is a CSV file separated by semicolons (“;”), whereas ELAN-Player requires a tab-separated CSV file (“\t”). The user interfaces of both annotation tools were presented to the study participants. The MaD-GUI’s video frame stands independently from the frame displaying inertial data, unlike the ELAN-Player’s, where the two are linked. This provides more flexibility in adjusting the sizes of each frame to suit the user’s preferences.

**NASA-TLX and Feedback Questionnaire:** The template and comprehensive results of the NASA-TLX questionnaire employed in our study are accessible for download from our GitHub repository. Additionally, we incorporated participant responses to two inquiries that solicited their opinions on the advantages and disadvantages of annotating IMU data using MaD-GUI or the ELAN-Player. A questionnaire was administered to participants after each annotation experiment, beginning with the WEAR dataset and then followed by the Wetlab dataset, as illustrated in Figure 1 in the main paper. This allowed us to gather feedback on their experience with each tool and identify any potential areas for improvement. The original NASA-TLX questionnaire (see [6]) comprises ten distinct categories to assess perceived workload. However, for our study, we employed a streamlined version of the NASA-TLX, encompassing six core dimensions: *Mental*

a. MaD-GUI displaying the Wetlab data



b. ELAN-Player displaying the WEAR data

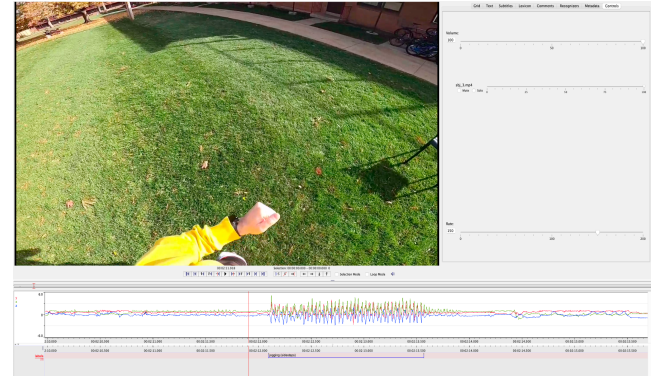


Fig. 1: (a) The MaD-GUI displaying the Wetlab data of participant 2. (b) The interface of the ELAN-Player displays the WEAR data of participant 3. Both tools display the video and inertial data simultaneously.

*Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration.* This abbreviated version was deemed sufficient for capturing the primary aspects of the annotation workload. Participants rated each dimension on a scale of 0 to 10, where 0 denoted the lowest rating and 10 the highest.

## REFERENCES

- [1] A. Hoelzemann, M. Bock, and K. Van Laerhoven, “Evaluation of video-assisted annotation of human imu data across expertise, datasets, and tools,” in *IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, 2024.
- [2] M. Ollenschläger, A. Küderle, W. Mehringer, A.-K. Seifer, J. Winkler, H. Gaßner, F. Kluge, and B. M. Eskofier, “MaD GUI: An Open-Source Python Package for Annotation and Analysis of Time-Series Data,” *MDPI Sensors*, vol. 22, no. 15, 2022.
- [3] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, “ELAN: A Professional Framework for Multimodality Research,” in *5th International Conference on Language Resources and Evaluation*, 2006.
- [4] M. Bock, H. Kuehne, K. Van Laerhoven, and M. Moeller, “WEAR: An Outdoor Sports Dataset for Wearable and Egocentric Activity Recognition,” *CoRR*, vol. abs/2304.05088, 2023.
- [5] P. M. Scholl, M. Wille, and K. Van Laerhoven, “Wearables in the Wet Lab: A Laboratory System for Capturing and Guiding Experiments,” in *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015.
- [6] S. G. Hart and L. E. Staveland, “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research,” in *Human Mental Workload* (P. A. Hancock and N. Meshkati, eds.), vol. 52 of *Advances in Psychology*, North-Holland, 1988.

<sup>2</sup><https://github.com/mad-lab-fau/mad-gui>

<sup>3</sup><https://mad-gui.readthedocs.io/>

<sup>4</sup><https://archive.mpi.nl/tla/elan>

<sup>5</sup><https://youtube.com/watch?v=cSFFSTUM4e0>

<sup>6</sup><https://youtube.com/watch?v=-GI5agFOPRM>

<sup>7</sup>[https://mpi.nl/tools/elan/docs/manual/Sec\\_The\\_Timeseries\\_Viewer.html](https://mpi.nl/tools/elan/docs/manual/Sec_The_Timeseries_Viewer.html)