Evaluation of Video-Assisted Annotation of Human IMU Data Across Expertise, Datasets, and Tools

1st Alexander Hoelzemann
Ubiquitous Computing
University of Siegen
Siegen, Germany
alexander.hoelzemann@uni-siegen.de

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

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

***Abstract*—Despite the simplicity of labels and extensive study protocols provided during data collection, the majority of researchers in sensor-based technologies tend to rely on annotations provided by a combination of field experts and researchers themselves. This paper presents a comprehensive study on the quality of annotations provided by expert versus novice annotators for inertial-based activity benchmark datasets. We consider multiple parameters such as the nature of the activities to be labeled, and the annotation tool, to quantify the annotation quality and time needed. 15 participants were tasked to annotate a total of 40 minutes of data from two publicly available benchmark datasets for inertial activity recognition, being simultaneously displayed both video and accelerometer data during annotation. We compare the resulting labels with the ground truth provided by the original dataset authors. Our participants annotated the data using two representative tools. Metrics like F1-Score and Cohen's Kappa showed experience did not ensure better labels. While experts were more accurate on the complex Wetlab dataset (51% vs 46%), Novices had 96% F1 on the simple WEAR dataset versus 92% for experts. Comparable Kappa scores (0.96 and 0.94 for WEAR, 0.53 and 0.59 for Wetlab) indicated similar quality for both groups, revealing differences in dataset complexity. Furthermore, experts annotated faster regardless of the tool. Given proven success across research, our findings suggest crowdsourcing wearable dataset annotation to non-experts warrants exploration as a valuable yet underinvestigated approach, up to a complexity level beyond which quality may suffer.**

***Index Terms*—Inertial sensors, Activity recognition, Sensor data annotation**

The first two authors contributed equally to this work.

I. Introduction and Related Work

The automatic recognition of activities through wearable inertial data has been identified as valuable information for numerous research fields and applications (see e.g., [1], [2], [3], [4]). The quality of benchmark datasets for such research notoriously depends on the richness of recorded sensor data, including the length of the recording, or the variety of participants performing the activities. The annotation of the data with activity class labels is often less prominent. While some annotation scenarios such as medical data can be restricted to field experts (e.g., medical staff), other application scenarios such as the recognition of activities of daily living require annotators to annotate a set of trivial labels (e.g., sitting, standing, etc.). Despite the simplicity of labels and extensive study protocols along with detailed activity descriptions, the majority of researchers in sensor-based technologies tend to rely on annotations provided by a combination of field experts and researchers themselves [5]. Along with most publicly available benchmark datasets failing to provide details on their annotation process, wearable-based data annotation remains to date a tedious, time-consuming task and requires researchers to dedicate a substantial time to it during data collection (up to 14 to 20 times longer than the actual recorded data, as for instance mentioned in [6]). Ultimately, the lack of documentation and absence of a common protocol for sensor-based data annotation has resulted in many published datasets following one's own individual and self-defined data annotation protocols. However, as evident from recent workshops dedicated to the annotation of datasets[1], and tutorials [2], there is scientific interest of the community in evaluating the impact, importance and applicability of annotation techniques, tools, and methods.

In other machine learning-related research areas such as Computer Vision [7], [8], [9], [10], a proven method for obtaining a substantial amount of annotations is the use of crowd-sourced annotation services such as Amazon Mechanical Turk [11], which are yet to be explored in the wearable-based community. Publicly available benchmark datasets that provide both image- and inertial-based sensor data remain scarce (see Table 1 in [12] for a curated list), yet authors of a majority of currently publicly-available datasets reported to rely primarily on video recordings to allow for an accurate offline labeling process [13], [12], [6], [14]. We present here a study that involves 10 novice and 5 expert annotators who labeled two, 20-minute data segments taken from two public inertial-based datasets used for activity recognition [12], [15] using two frequently used open-source visualization tools [16], [17]. The primary objective of this publication is to address the following set of research questions:

1) To what extent are novices - annotators with limited expertise on inertial sensor data - capable of annotating such activity data using presently available open-source software?

[1] https://text2hbm.org/arduous/

[2] https://www.ubicomp.org/ubicomp-iswc-2023/program/tutorials/

2) Does the prior acquaintance of annotators with sensor-based data significantly influence resulting annotations?
3) What are the obstacles that must be overcome and what limitations exist in commonly employed tools and procedures?

A. Data Annotation Challenges

With dedicated calls for datasets at conferences or journals, such as NeurIPS[3] or IMWUT[4] the scientific community acknowledges the need to have a broad spectrum of benchmark datasets. However, with the rising interest in publishing new datasets, we face complex challenges regarding the labeling process itself. According to Yordanova [18], researchers working on wearable, sensor-based datasets face 4 challenges while annotating data. These challenges are: (1) The gap of knowledge between system designers and knowledge experts, (2) the process of label engineering, (3) defining the meaning of a label, and (4) annotating big data.

(1) The gap of knowledge between system designers and knowledge experts describes the problem that in many scientific projects the people who design the classifying algorithm and the people who label the recorded dataset are not the same. This problem can lead to a missing common language between both groups and finally to a bias in the dataset caused by simplified or even naive ground truth or the missing ability of the domain expert to define a formal definition for a complex activity. Moreover, data work is often carried out by annotators who are not co-located in the same geography or culture as the ML practitioners which can serve to further distance data labor from its outputs [19].

(2) The process of label engineering refers to the fact that many commonly used datasets for sensor-based human activity recognition, such as those summarized in Table 1 (Hoelzemann *et al.* [13]), lack transparency regarding how they were collected, reviewed, and labeled. This makes it difficult to thoroughly evaluate the accuracy and correctness of the labeling. Omitting this information obscures potential flaws or biases that may exist in the datasets. For instance, research publications presenting new datasets often do not provide enough details about the decisions undertaken during data collection, curation, and annotation [20], [21], [22].

(3) Defining the meaning of a label can be a challenging task for a researcher, depending on the complexity of the activity classes of a dataset. Atomic labels, such as *sit, stand, walk, etc.* are unambiguous, however more complex labels like *sit_to_stand, walk_to_run, doing_laundry, cleaning_dishes, etc.* can be ambiguous for both, the annotator, as well as the learning algorithm. Since scientists are looking for more complex classes and more challenging datasets, this problem needs to be taken into account by researchers through developing tools that can help with defining such structured and complex labels.

(4) Annotating big data or datasets, outlines the need for annotation tools that can help with the annotation of data in large quantities by multiple people. Whenever datasets are recorded, especially in-the-wild, researchers tend to employ self-annotation and/or in-situ annotation methods [5]. These methods have the advantage that the workload for an annotator is moved to the participant itself, however, they are susceptible to producing an incomplete ground truth [18], [23]. Participants tend to forget to annotate their data or use synonyms for the same activity class [18]. Employing deep learning methodologies like transfer learning and active learning can partially address the labeling bottleneck, but human verification is still required to check and correct the automatically generated labels.

B. Activity Recognition Annotation Challenges

In theory, the video frames and IMU sensor samples captured between two key synchronization points should be properly aligned timewise. However, in practice, synchronizing the video and IMU streams precisely can pose difficulties. Challenges arise from clock drift between the camera and IMU devices, as well as different sampling rates for video frames versus IMU data samples [12]. Even with initial synchronization (e.g. via synchronization jumps), these issues can cause the alignment between video frames and IMU samples to drift over time, which makes it challenging to guarantee that every individual video frame and corresponding IMU data sample remain correctly synchronized throughout the captured data sequences between key synchronization points. Moreover, as Bulling *et al.* [24] note, multiple sensors may be subject to sensor drift. That is, the sampling rate of various sensors in a system may deviate from their original calibration over time, resulting in inconsistent measurements across sensors. Careful engineering and calibration of the data capture setup is required to minimize time drift and maintain tight synchronization between the video and IMU modalities as well as multiple sensors.

To address the challenges of synchronizing and annotating multimodal activity data, researchers have developed several specialized tools, for instance, the MaD-GUI [17], ELAN-Player [16], the Wearable Development Kit (WDK) [25], and Signaligner [26]. We decided to include MaD-GUI due to its accessibility as a Python package and its easy adaptability, as well as ELAN-Player due to its prevalent use as an annotation tool within our research community.

II. Study Design

Our study utilized a convenience sample of 15 volunteers. 5 participants with experience in activity recognition were classified as experts, while 10 participants without prior experience in activity recognition nor inertial dataset labeling were deemed novices, representative of crowd-sourced annotation services. The WEAR [12] dataset contains fitness activities recorded during untrimmed outdoor workout sessions. Video footage was recorded using a head-worn camera capturing the egocentric view of each participant. In contrast, the Wetlab [15] dataset had participants perform lab experiments being

[3] https://neurips.cc/Conferences/2023/CallForDatasetsBenchmarks
[4] https://dl.acm.org/journal/imwut/dataset-papers-guidelines

recorded from a top-down, stationary perspective. Comparing the two datasets, labeled activities in WEAR are more sequential and block-wise, whereas activities contained in the Wetlab dataset, given its specific use case, are more ambiguous to third-party observers. Additionally, Wetlab exhibits less intense acceleration, posing challenges in identifying actions. Each participant annotated 20 minutes of data from the WEAR dataset and 20 minutes from the Wetlab dataset. Regardless of the used annotation tool (ELAN or MaD-GUI), each participant was simultaneously displayed both the video stream along with accelerometer data captured at the right-wrist. To ensure the quality and consistency of annotations, we restricted the study to a one-hour duration. This decision was influenced by previous research showing prolonged annotation tasks can lead to fatigue [27] and boredom [28], reducing quality. Moreover, the complexity of sensor datasets can make it hard to maintain focus during extended sessions [29]. By dividing annotation into shorter sessions, annotators can take breaks and recharge, potentially improving quality.

Randomization. We divided both datasets into two 10-minute segments and randomly assigned either the MaD-GUI or ELAN-Player first to each participant. The study procedure (Figure 1) then involved a briefing to introduce the tool's functions and activity classes in the subset. Participants annotated the WEAR dataset first, followed by the Wetlab dataset. This order was chosen because of the comparatively lower mental demand of the WEAR dataset, comprised primarily of long activity blocks. In total, each participant annotated 40 minutes of data by completing the procedure with both tools.

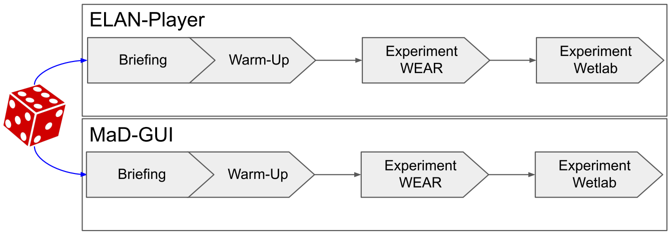


Fig. 1. The study protocol consists of four main components: (1) Briefing, (2) (Tool) Warm-Up, (3) Experiment WEAR, and (4) Experiment Wetlab. Randomization, indicated by the dice, was used to determine the participant's initial tool to avoid learning effects.

Briefing and Warmup. When first introduced to the ELAN-Player or MaD-GUI, participants took part in a briefing on the tools and labeling classes. During the briefing phase, the supervising researcher explained the annotation procedure, which differs for each tool, including how to add, delete, or modify annotations. Furthermore, each activity class was explained in detail along with a short demonstration. The briefing was followed by a 5-minute warm-up session. During this warm-up, participants became familiar with the tool and dataset classes while watching a short segment from a different dataset participant.

Class Description. The first 20 minutes of data were used from the third WEAR subject and the second Wetlab subject. Neither subject was shown during the warm-up. The WEAR subset consisted of six, and the Wetlab subset of seven classes:

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

For the Wetlab dataset, an additional *mixing* class was included in the label options, though not part of the original label set. This complex, multi-step activity ("mix into 200ml beaker, add 1ml detergent, and stir" [15]) was deliberately added to challenge participants and introduce ambiguity in identifying and labeling activities. During the study, participants were left alone in a room without time constraints. After annotating 10 minutes of data from both datasets, participants were requested to complete a questionnaire. The questionnaire included the NASA Task Load Index (*NASA-TLX*) assessment [30], [31], which aims to evaluate the perceived workload during the annotation task. Furthermore, participants were provided with two open text fields to express their positive and negative experiences, as well as their opinions about the tool. The test distinguishes 6 different evaluation categories (Mental, Physical and Temporal Demand, Performance, Effort and Frustration). Each category is rated between 0 and 100 with a 5-point step size.

Evaluation Metrics. Each participant's annotation session was assessed using four metrics. First, we measured the *time* to annotate each 10-minute segment. Second, we calculated the *Cohen's Kappa Score* [32] between annotations and ground truth to measure inter-annotator agreement. The *Cohen's Kappa Score or Cohen-*κ is calculated as $\kappa = (p_0 - p_e)/(1 - p_e)$ with p_0 being the relative observed agreement between the annotator and the ground truth, and p_e being the hypothetical probability of chance agreement. Third, we computed the *F1-score* [33] which is the harmonic mean between the precision and recall score of a participant's annotations compared with the ground truth labels. Finally, in order to assess the overall ability of an annotator to spot action segments within the data stream, we calculate the *NULL-class accuracy* score.

III. Results

Figure 2 provides a color-coded illustration of provided annotations of each participant split across each data segment and involved dataset. Table I depicts times measured by the study supervisor that each participant needed to finish each subset-specific annotation task, as well as the corresponding F1-score, Cohen-κ value, and NULL-accuracy calculated on the annotated data. One can see that the accuracy and consistency of provided annotations varied heavily depending on which dataset was to be annotated by participants. Comparing these calculated evaluation metrics, it is evident that the WEAR dataset is, due to its more structured sequences in activities, less challenging for any annotator. The high overall F1-score of 95.03% and an average Cohen-κ of 0.96 indicates that almost every participant was capable of annotating the data consistent to the ground truth annotated by the original authors. Note that confusion amongst labels (see subject 2 in

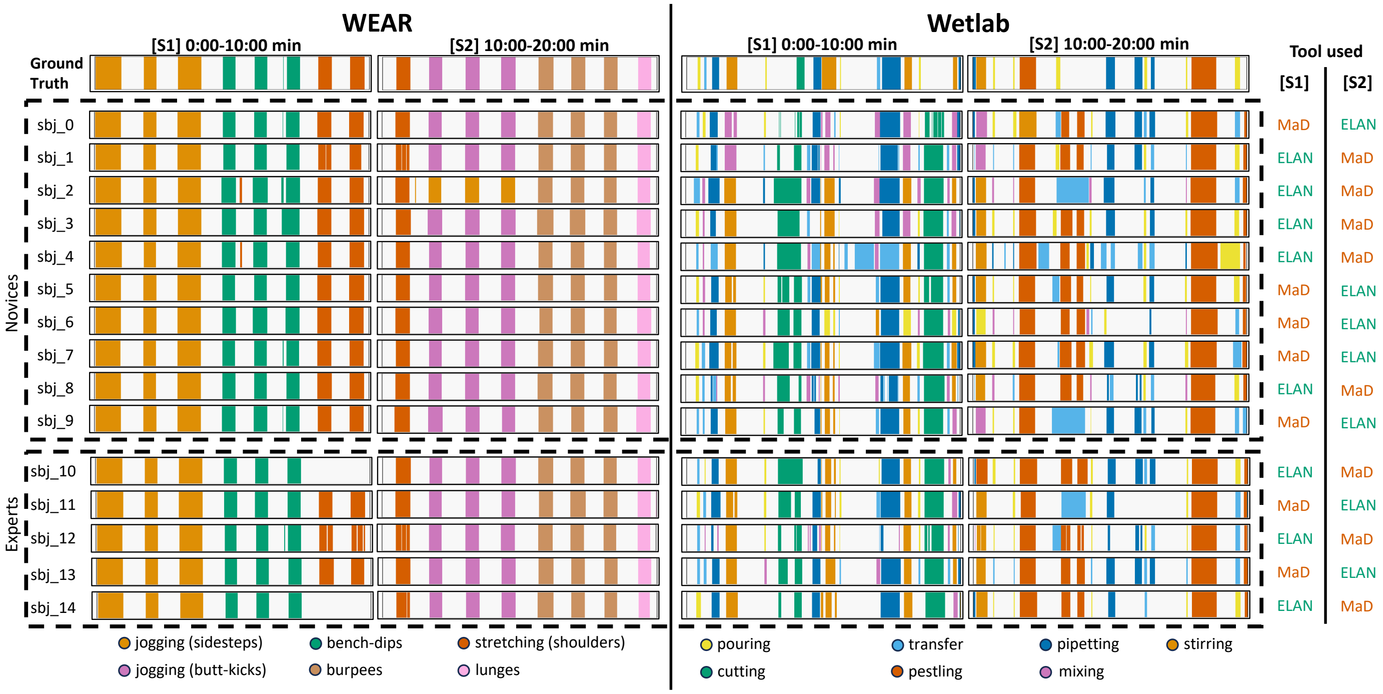


Fig. 2. Figure 2 shows color-coded annotations from all participants, categorized by expertise (expert or novice), dataset (WEAR or Wetlab), and 10-minute data segment (S1 or S2). The top row displays the original authors' ground truth annotations for each segment. The right column indicates the annotation tool used for specific segments, assigned randomly as depicted in Figure 1. The color legend below the annotations denotes the label classes occurring in the datasets. In summary, this figure provides a visual overview of the collected annotations across expertise levels, datasets, segments, and tools. The color-coding allows comparison to the ground truth and highlights labeling patterns.

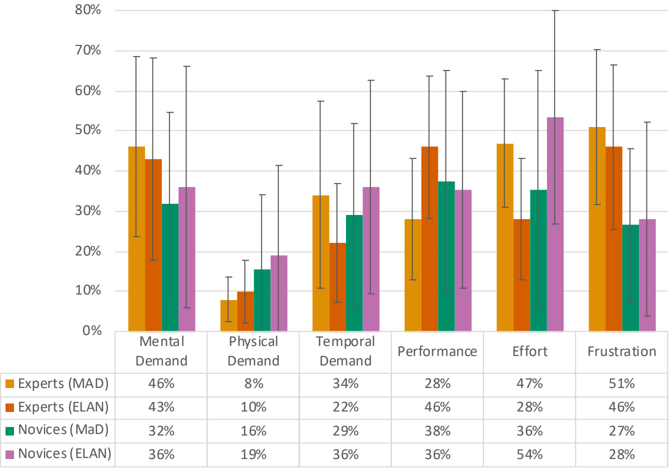
TABLE I
SUMMARY OF F1-SCORES, COHEN-κ VALUES, NULL-ACCURACIES, AND INDIVIDUAL TIMES NEEDED TO ANNOTATE THE SUBSETS OF THE WEAR AND WETLAB DATASET USING EITHER THE MAD-GUI OR ELAN-PLAYER. COLORS INDICATE 🟩 NOVICES, 🟩 NOVICES AVERAGE AND STANDARD DEVIATION, 🟦 EXPERTS, 🟦 EXPERTS AVERAGE AND STANDARD DEVIATION, 🟥 OVERALL AVERAGE AND STANDARD DEVIATION.

	WEAR					Wetlab				
				Annotation Time					Annotation Time	
Subject	F1-score	Cohens-κ	NULL-accuracy	MaD	ELAN	F1-score	Cohens-κ	NULL-accuracy	MaD	ELAN
sbj_0	98.96%	0.9837	99.00%	15:45	8:32	37.62%	0.5323	84.33%	29:42	16:25
sbj_1	97.31%	0.9666	99.81%	18:08	20:30	39.92%	0.5397	83.11%	22:56	20:34
sbj_2	80.73%	0.8375	96.81%	23:10	17:12	43.49%	0.5385	69.82%	22:20	32:43
sbj_3	96.95%	0.9549	95.45%	12:13	8:47	60.53%	0.5560	77.39%	15:46	18:34
sbj_4	97.31%	0.9637	98.23%	31:30	21:15	35.56%	0.3419	60.78%	47:30	71:00
sbj_5	98.82%	0.9829	99.42%	19:46	12:42	60.57%	0.6386	79.23%	22:34	15:51
sbj_6	97.64%	0.9639	98.13%	29:43	16:35	29.93%	0.4512	79.55%	31:30	18:04
sbj_7	98.99%	0.9845	99.64%	25:41	17:18	62.20%	0.5827	71.24%	35:40	28:49
sbj_8	98.99%	0.9854	99.52%	17:02	29:26	44.85%	0.4944	78.60%	24:45	53:19
sbj_9	97.43%	0.9640	97.98%	16:35	17:54	48.38%	0.5912	80.12%	21:34	19:58
Avg.	96.31 ± 5.54%	0.9587 ± 0.044	98.40 ± 1.40%	**20:57 ±6:21**	**17:01 ±6:11**	46.31 ± 11.42%	0.5267 ± 0.0829	76.42 ± 7.14%	**27:26 ±9:04**	**29:32±18:30**
sbj_10	84.82%	0.8879	98.90%	6:59	11:30	47.71%	0.5844	77.99%	11:45	16:50
sbj_11	97.71%	0.9654	98.07%	13:45	5:12	58.48%	0.5806	78.44%	16:17	7:57
sbj_12	96.96%	0.9627	99.72%	12:53	10:17	43.82%	0.5087	83.74%	12:50	16:08
sbj_13	98.53%	0.9811	99.42%	10:00	10:17	50.77%	0.6319	80.63%	20:19	16:08
sbj_14	84.29%	0.8863	99.67%	9:24	8:57	54.53%	0.6381	84.38%	14:32	19:30
Avg.	92.46% ± 7.24	0.9367 ± 0.0458	99.16% ± 0.69	**10:36 ±2:44**	**9:15 ±2:26**	51.06% ± 5.72	0.5888 ± 0.0519	81.04% ± 2.95	**15:09 ±3:22**	**15:19±4:21**
Avg.	95.03% ± 6.18	0.9514 ± 0.0443	98.65% ± 1.24	**17:30 ±7:19**	**14:26 ±6:22**	47.89% ± 9.92	0.5474 ± 0.0782	77.96% ± 6.35	**23:20 ±9:36**	**24:47± 16:32**

Figure 2) is probably to be classified as inadvertent mistakes, for instance by selecting the wrong label within a selection window.

The NULL-accuracy of 98.65% further suggests that every participant distinguished successfully relevant action segments in the data with only two subjects 10 and 14 being the exception, which were not able to detect and label the activity ● stretching (shoulders) within the first ten minutes of data. As both annotators were classified as experts, the missed annotations not being characterized by peaks in the acceleration datastream and the annotation taking place at the very end of the dataset segment, failing to annotate said activity might be caused by said experts relying too much on their own expertise in recognizing action segments in the inertial data without feeling the need to cross-check them with the corresponding video stream. Contrarily, the evaluation metrics obtained on the Wetlab dataset make it evident that experts are on average able to score a higher F1-score, Cohen-κ, and NULL-accuracy compared to novices, with respective average improvements being around 4.75%, 0.0621, and 4.62%. Figure 2 further reveals that for the Wetlab dataset the activities ● pouring, ● transfer, ● stirring and ● mixing were frequently confused with each other by both experts and novices. While the WEAR dataset has been homogeneously annotated by (almost) every participant, the Wetlab dataset covers a lesser-known application scenario of activities in a DNA-extraction experiment, causing participants to lack a deeper understanding of the involved labels and their semantics, regardless of their experience level in annotating inertial-based data. In summary, it is noticeable that experts are about twice as fast as novices while achieving high F1-scores, Cohen-κ values, and NULL-accuracies, regardless of the tool or dataset they were tasked to use. As mentioned earlier, the Wetlab dataset comprises of short, complex, and interconnected activities, making it challenging for annotators that lack expert domain knowledge to differentiate between them. Annotating these subsets proved to be more demanding for especially novice participants, particularly due to the introduction of the intricate task of ”mixing”. Interestingly, though annotation quality varied to a larger degree amongst novice annotators, it is noteworthy that novice subjects 3, 5, and 7 achieved the highest overall annotation performance on the Wetlab dataset.

Cognitive Workload: Comparing results obtained from the *NASA-TLX* assessments, one can see large deviations amongst participants (see Figure 3). In general, the results of both expertise groups indicate that the cognitive workload and temporal demand were perceived to be low to mediocre for both tools. Further, frustration was higher with expert annotators while physical demand was perceived higher with novice annotators.



	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration
■ Experts (MAD)	46%	8%	34%	28%	47%	51%
■ Experts (ELAN)	43%	10%	22%	46%	28%	46%
■ Novices (MaD)	32%	16%	29%	38%	36%	27%
■ Novices (ELAN)	36%	19%	36%	36%	54%	28%

Fig. 3. The results of the NASA-TLX [30] assessment indicate that the perceived cognitive workload remains consistent regardless of the tool used. However, participants reported slightly higher levels of Mental Demand, Physical Demand, Performance, and Effort when using the ELAN-Player, while perceiving lower levels of Frustration. Temporal Demand was considered as equal for both tools.

Overall, the ELAN-Player was regarded as the slightly more suitable tool: Both groups felt more confident in annotating samples correctly using the ELAN-Player and were less frustrated by its user interface. In general, participants who started with the ELAN-Player often expected more functionality with regards to the control options of the MaD-GUI video player, since the ELAN-Player (first released in 2006) is a well-established tool that offers additional video controls, such as fast forward/ slow motion or forwarding/ reversing the video in fine-granular steps. A full list of comments about the tools is included in this paper's supporting material document.

IV. Conclusions

This paper presented a quantitative assessment of the difference in the quality of annotations provided by expert and novice annotators in the context of labeling wearable inertial data. The study involved participants annotating 40 minutes of data: two 10-minute segments originating from two publicly available inertial datasets used for wearable activity recognition [15], [12]. Participants used two well-known and open-source visualization tools [16], [17] which are capable of displaying both inertial and video data of the respective datasets simultaneously. Evaluation metrics, i.e. F1-score, Cohen-κ, and NULL-accuracy showed that participants classified as experts, who have experience in annotating inertial data and can benefit from prior experience with inertial sensors, were able to annotate the data segments quicker and overall more consistent in quality. Further, depending on which dataset was to be labeled, annotations varied significantly in quality. While the WEAR dataset, which contains mostly easy-to-understand fitness exercise labels, was labeled almost completely in line with the ground truth, the Wetlab dataset, which consists of a more complex application scenario and consists of labels with higher ambiguity, showed higher confusion amongst all annotators. Although confusion among novice annotators was higher, a significant amount of novices were able to provide annotations of the same quality as compared to experts, with some even outperforming experts by a significant margin (approx. 2-4% F1-Score).

We identify four key takeaways of our study:

1) Both experts and novices encountered similar issues. We thus believe that annotation of wearable activity recognition data should not to be considered exclusive to individuals experienced with wearable technologies.
2) High NULL-accuracies along with a low F1-scores of activity classes, suggest that correctly conveying semantics and overall understanding of the segments to be labeled remains the main hurdle to maximize the likelihood of consistency amongst annotators.
3) Agreeing results for Cohen's Kappa indicate that variations in dataset characteristics, such as data and class complexity, could be the determining factors in identifying datasets that are more suitable to be annotated by novices than others.
4) Given that recording plans for inertial benchmark datasets often contain detailed explanations of performed activities, we argue many publicly available datasets could be crowdsourced for annotation purposes with proper measures in place.

What remains to be investigated is which measures, such as sample video clips of action segments, prove to be most effective in conveying the semantics of activities to these novice annotators. Considering its proven success across numerous research domains, our findings imply that crowdsourcing wearable dataset annotation to non-expert annotators merits further exploration as a valuable yet under-investigated approach, up to a given level of data complexity beyond which the label quality may suffer.

Funding and Ethics

This project is funded by the House of Young Talents at the University of Siegen and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the project WASEDO (506589320) as well as the project PervaSafe Computing (425868829), which is part of Priority Program SPP2199 Scalable Interaction Paradigms for Pervasive Computing Environments. The studies involving human participants were reviewed and approved by the Ethics Committee of the University of Siegen (ethics vote #ER_12_2019). The participants provided their written informed consent to participate in this study and were briefed prior to the data capture about the study goals.

References

[1] L. Bao and S. S. Intille, "Activity Recognition From User-Annotated Acceleration Data," *Pervasive Computing*, 2004.

[2] W. W. Tryon, "Activity Measurement," in *Clinician's handbook of adult behavioral assessment*, Practical resources for the mental health professional, Academic Press, 2006.

[3] D. J. Patterson, D. Fox, H. Kautz, and M. Philipose, "Fine-Grained Activity Recognition by Aggregating Abstract Object Usage," in *Ninth IEEE International Symposium on Wearable Computers*, 2005.

[4] J. A. Ward, P. Lukowicz, G. Tröster, and T. E. Starner, "Activity Recognition of Assembly Tasks Using Body-Worn Microphones and Accelerometers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[5] A. Hoelzemann and K. Van Laerhoven, "A Matter of Annotation: An Empirical Study on in Situ and Self-Recall Activity Annotations From Wearable Sensors," *CoRR*, vol. abs/2305.08752, 2023.

[6] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. d. R. Millàn, "Collecting Complex Activity Datasets in Highly Rich Networked Sensor Environments," in *IEEE Seventh International Conference on Networked Sensing Systems*, 2010.

[7] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The Open Images Dataset V4," *International Journal of Computer Vision*, vol. 128, no. 7, 2020.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), 2014.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-scale Hierarchical Image Database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[11] A. Sorokin and D. Forsyth, "Utility Data Annotation With Amazon Mechanical Turk," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2008.

[12] 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.

[13] A. Hoelzemann, J. L. Romero, M. Bock, K. Van Laerhoven, and Q. Lv, "Hang-Time HAR: A Benchmark Dataset for Basketball Activity Recognition Using Wrist-Worn Inertial Sensors," *MDPI Sensors*, vol. 23, no. 13, 2023.

[14] P. Zappi, C. Lombriser, T. Stiefmeier, E. Farella, D. Roggen, L. Benini, and G. Tröster, "Activity Recognition From On-Body Sensors: Accuracy-Power Trade-Off by Dynamic Sensor Selection," in *European Conference on Wireless Sensor Networks*, 2008.

[15] 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.

[16] 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.

[17] 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.

[18] K. Yordanova, "Challenges providing ground truth for pervasive healthcare systems," *IEEE Pervasive Computing*, vol. 18, no. 2, 2019.

[19] M. L. Gray and S. Suri, *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books, 2019.

[20] E. Denton, A. Hanna, R. Amironesei, A. Smart, and H. Nicole, "On the genealogy of machine learning datasets: A critical history of ImageNet," *Big Data & Society*, vol. 8, no. 2, 2021.

[21] M. K. Scheuerman, K. Wade, C. Lustig, and J. R. Brubaker, "How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis," *ACM on Human-Computer Interaction*, vol. 4, no. CSCW1, 2020.

[22] M. K. Scheuerman, A. Hanna, and E. Denton, "Do datasets have politics? Disciplinary values in computer vision dataset development," *ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, 2021.

[23] K. Van Laerhoven, D. Kilian, and B. Schiele, "Using rhythm awareness in long-term activity recognition," in *12th IEEE International Symposium on Wearable Computers*, 2008.

[24] A. Bulling, U. Blanke, and B. Schiele, "A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors," *ACM Computing Surveys*, vol. 46, no. 3, 2014.

[25] J. Haladjian, "The wearables development toolkit: An integrated development environment for activity recognition applications," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 4, 2020.

[26] A. Ponnada, S. Cooper, Q. Tang, B. Thapa-Chhetry, J. A. Miller, D. John, and S. Intille, "Signaligner Pro: A tool to explore and annotate multi-day raw accelerometer data," in *IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events*, 2021.

[27] T. Stoev, K. Yordanova, and E. L. Tonkin, "Experiencing Annotation: Emotion, Motivation and Bias in Annotation Tasks," in *IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events*, 2023.

[28] M. Meier, C. Martarelli, and W. Wolff, "Bored Participants, Biased Data? How Boredom Can Influence Behavioral Science Research and What We Can Do About It," *PsyArXiv*, 2023.

[29] Y. Liu, L. Nie, L. Han, L. Zhang, and D. S. Rosenblum, "Action2Activity: Recognizing Complex Activities From Sensor Data," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[30] N. N. Aeronautics and S. Administration)., "NASA Task Load Index (NASA-TLX), Version 1.0: Paper and Pencil Package," 1986.

[31] 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.

[32] R. Artstein and M. Poesio, "Inter-Coder Agreement for Computational Linguistics," *Computational Linguistics*, vol. 34, no. 4, 2008.

[33] N. Chinchor and B. M. Sundheim, "MUC-5 evaluation metrics," in *Fifth Conference on Message Understanding*, 1993.