Evaluating the Accuracy of Automated Data Annotation

Sahil Jagtap

May 28, 2025

Abstract

This report provides an in-depth evaluation of an automated data annotation tool developed for labeling sensor data. Using a manually annotated dataset as ground truth, the tool's output was assessed for accuracy. The results show an overall labeling accuracy of 27.26%, with significant variance across label classes. The report investigates sources of error, limitations in the current logic, and directions for future improvement.

1 Introduction

Automating annotation is critical for scaling supervised learning and behavior recognition systems, especially when working with time-series accelerometer data. Manual annotation, while accurate, is time-consuming and subjective. This project aims to automate the annotation process with logic-based label assignment, followed by validation against expert-labeled ground truth.

2 Data and Methodology

The dataset consisted of CSV files containing sensor readings along with human-provided labels. Each record includes:

- `unixTimestampInMs`
- Accelerometer readings: `x`, `y`, `z`
- `readableTime`
- `label`

Two files were used for evaluation:

1. Manually labeled CSV: `P19_SJ_manual.csv`
2. Automatically labeled CSV: `P19_automated.csv`

The evaluation script matched rows using timestamp and acceleration data, comparing labels to compute:

- Total correct matches
- Total comparisons
- Overall accuracy
- Per-label accuracy breakdown

3 Results

- Total comparisons: 63,745
- Correct matches: 17,378
- Overall accuracy: **27.26%**

Per-label Accuracy Breakdown (excerpt)

```
Discard:      93.12% (e.g., easy to identify based on data voids)
On Task:      41.89%
Off Task:     11.44%
```

4 Analysis and Discussion

The low overall accuracy can be attributed to:

1. **Hardcoded Heuristics:** The logic in `annotator.py`, `detector.py`, and `process_annotations.` appears to use fixed rule-based thresholds, which are brittle to noise and inter-participant variation.
2. **Lack of Contextual Features:** Temporal continuity or multi-feature fusion (like variance over time windows) is not leveraged.
3. **Left-Handed Subjects:** Subjects P9 and P14 are left-dominant, but the annotation logic does not seem to adjust orientation or region accordingly.
4. **Misaligned Labeling Semantics:** Discrepancies in how ”Off Task” or ”On Task” is defined manually vs. programmatically may lead to mismatches even when the motion pattern is similar.

5 Recommendations

- Integrate sliding window features (e.g., mean, variance) to capture temporal motion trends.
- Parameterize dominant hand and allow for region flipping.
- Introduce a probabilistic labeling model or train a lightweight classifier using partial labeled data.
- Add visual inspection and debugging tools to see mislabeled segments.

6 Conclusion

While the current automation pipeline lays the groundwork, it falls short in reliably matching human annotations. Enhancing the system with contextual analysis and adaptive logic can significantly improve performance.

Acknowledgements

Thanks to Anika for providing guidance, recordings, and manual annotations.