ChildLens: An Egocentric Video Dataset for Activity Analysis in Children

Nele-Pauline Suffo[1], Pierre-Etienne Martin[2], Daniel Haun[2], & Manuel Bohn[1, 2]

[1] Institute of Psychology in Education, Leuphana University Lüneburg

[2] Max Planck Institute for Evolutionary Anthropology

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line. Enter author note here.

The authors made the following contributions. Nele-Pauline Suffo: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Manuel Bohn: Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to Nele-Pauline Suffo, Universitätsallee 1, 21335 Lüneburg. E-mail: nele.suffo@leuphana.de

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline. Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines. One sentence clearly stating the **general problem** being addressed by this particular study. One sentence summarizing the main result (with the words "**here we show**" or their equivalent). Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge. One or two sentences to put the results into a more **general context**. Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

ChildLens: An Egocentric Video Dataset for Activity Analysis in Children

Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed ac purus sit amet nisl tincidunt tincidunt. Nullam nec turpis at libero tincidunt tincidunt. Sed nec mi nec nunc tincidunt tincidunt. Nullam nec turpis at libero tincidunt tincidunt. Sed nec mi nec nunc

Dataset Overview

Activity Classes. The ChildLens dataset contains a total of 14 activity and 5 location classes. The activities are based on the actions of the child in the video and can be divided into *person-only* activities, such as "child talking" or "other person talking, and *person-object interaction* activities, such as"drawing" or "playing with object". You can find the complete list of activity classes in the appendix.. The activities can be further divided into *audio-based*, *visual-based*, and *audio-visual* activities. The following list provides an overview of the different activity types:

- **Audio-based activities**: *child talking*, *other person talking*, *overheard speech*, *singing / humming*, *listening to music / audiobook*
- **Visual-based activities**: *watching something*, *drawing*, *crafting things*, *dancing*
- **Multimodal activities**: *playing with object*, *playing without object*, *pretend play*, *reading book*, *making music*

The location classes describe the current location of the child in the video and include *livingroom*, *playroom*, *bathroom*, *hallway*, and *other*.

Statistics. We have varying numbers of clips for each of the 14 activity classes, ranging from x to x clips per class. The duration of the clips differs depending on the activity; for example, audio-related actions like "child talking" may only last a few seconds,

Table 1

Number of clips per class

training	validation	testing
10	10	10

while activities like “reading a book” may last several minutes. The xxx video files are divided into *xx-xx* training videos, *xx** validation videos, and *xx* testing videos for each class.

Table 1 provides an overview of the number of clips per class in the training, validation, and testing sets.

Exhaustive multi-label annotations. The dataset provides detailed annotations for each video file. These include the start and end times of each activity, the activity class, and the child’s current location within the video. If multiple activities occur simultaneously in a video, each activity is individually labeled and extracted as a separate clip. For example, if a segment of a video shows a child “reading a book” while also “talking,” two separate clips are created: one for “reading a book” and another for “child talking.”

How the Dataset was Built

Step 1: Generating a labeling strategy

Step 2: Manual labelling process

Discussion: Dataset bias

Benchmark Performance

Boundary-Matching Network

We utilize the BMN model (Lin, Liu, Li, Ding, & Wen, 2019) for temporal activity localization.

VTC

Implementation details

Conclusion

Results

Discussion

References

We used R [Version 4.4.1; R Core Team (2024)] for all our analyses.

Lin, T., Liu, X., Li, X., Ding, E., & Wen, S. (2019). *BMN: Boundary-Matching Network for Temporal Action Proposal Generation*. arXiv. https://doi.org/10.48550/ARXIV.1907.09702

R Core Team. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Appendix

List of ChildLens Activity Classes

The dataset contains the following list of activities. The number of clips for each activity class is indicated by the number in brackets behind each class.

1. playing with object TBD
2. playing without object TBD
3. pretend play TBD
4. watching something TBD
5. reading book TBD
6. child talking TBD
7. other person talking TBD
8. overheard speech TBD
9. drawing TBD
10. crafting things TBD
11. singing / humming TBD
12. making music TBD
13. dancing TBD
14. listening to music / audiobook TBD

List of ChildLens Location Classes

1. livingroom
2. playroom
3. bathroom
4. hallawy
5. other