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

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 *multimodal* 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 clips are divided into *xx-xx* training clips, *xx** validation clips, and *xx* testing clips for each class. Table 1 provides an overview of the numbers.

Exhaustive multi-label annotations. The dataset provides detailed annotations for each video file. These annotations specify the child’s current location within the video, the start and end times of each activity, the activity class, and whether the child is engaged alone or with somebody else. For every person involved in the activity, we capture age and gender. If multiple activities occur simultaneously in a video, each activity is individually labeled and extracted as a separate clip. For example, if a segment shows a child “reading a book” while also “talking,” two separate clips are created: one for “reading a book” and another for “child talking.” This exhaustive labeling strategy ensures that each activity is accurately represented in the dataset.

How the Dataset was Built

This section outlines the steps taken to create the ChildLens dataset. We provide detailed information on the video collection process, the labeling strategy employed, and the generation of activity labels.

Step 1: Collection of Egocentric Videos

The ChildLens dataset consists of egocentric videos recorded by children aged 3 to 5 years. A total of xx children from families in Leipzig, Germany, participated in the study.

The videos were captured at home using a camera embedded in a vest worn by the children. This setup allowed the children to move freely throughout their homes while recording their activities. The camera captured within the child's field of view, while an integrated camera recorded the audio. Additionally, the parents were handed a small checklist of activities to record, ensuring that a variety of activities were captured in the videos. The focus was on everyday activities that children typically engage in, such as playing with toys, reading books, or drawing. The videos were recorded over a period of xx months, resulting in a total of xx hours of video footage.

Step 2: Creation of Labeling Strategy

The labeling strategy for the ChildLens dataset was designed to capture the child's daily activities accurately and is based on the instructions given to the parents. The parents were provided with a list of activities to record, including the following activities: - Child is invited to read a book together with an adult - Child is invited to play with toys alone - Child is invited to play with toys with someone else (adult or child) - Child is invited to draw/craft something

After an initial review of the videos, we identified the most common activities that children engage in throughout the day. The activity classes in the dataset are derived from these activities. We chose to differentiate, for example, between "drawing" and "crafting things" or "making music" and "singing/humming" in order to make the activities more granular. We also added the concepts of "overheard speech", which describes situations in which the child is not directly involved in a conversation but can hear it, and "pretend play", which refers to when the child is engaged in imaginative play. This approach allowed us to capture the diversity of activities that children engage in and create a comprehensive dataset for activity analysis.

Step 3: Manual Labeling Process

The videos were manually annotated by native German speakers who watched each video and labeled the activities present in the footage. The annotators marked the start and end points of each activity, ensuring that the annotations were accurate and detailed. The labeling process was conducted using the SuperAnnotate platform, which allowed for efficient annotation and review of the videos.

- labeling process means looking at the whole video and mark start and end point of each activity present in the list of activities
- we annotate the videos using humans
- native germans for labeling task
- use SuperAnnotate platform for labeling
- review system: always two annotators label the same video, *-TODO: ask how exactly the review process works
- in setup meeting we discussed the labeling strategy and the classes and handed over the annotation strategy to the annotators
- before starting annotation process, we made sure that all questions were answered and that the annotators understood the task
- before the actual annotation process started, the annotators labeled 25 videos to get a feeling for the task
- the annotations then were reviewed by us and feedback was shared
- in total three feedback loops were conducted to ensure that the annotators followed the labeling strategy

Benchmark Performance

Implementation details

Boundary-Matching Network

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

VTC

For the visual-based activities, we use the Voice Type Classifier (Lavechin, Bousbib, Bredin, Dupoux, & Cristia, 2020).

Conclusion

Discussion

Dataset bias

General Discussion

References

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

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). *An open-source voice type classifier for child-centered daylong recordings*. arXiv. https://doi.org/10.48550/ARXIV.2005.12656

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