

ChildLens: An Egocentric Video Dataset for Activity Analysis in Children

Nele-Pauline Suffo<sup>1</sup>, Pierre-Etienne Martin<sup>2</sup>, Daniel Haun<sup>2</sup>, & Manuel Bohn<sup>1, 2</sup>

<sup>1</sup> Institute of Psychology in Education, Leuphana University Lüneburg

<sup>2</sup> 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* activities, such as “drawing” or “playing with object”. You can find the complete list of activity classes with a brief description in the appendix. The activities can be further divided into *audio-based*, *visual-based*, and *multimodal* activities, as presented in figure 1. 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.** The ChildLens dataset comprises 106.10 hours of video material recorded by 61 children aged 3 to 5 years. For each child, the duration of recorded video material varies between 4.03 and 303.42 minutes. A detailed distribution of the video

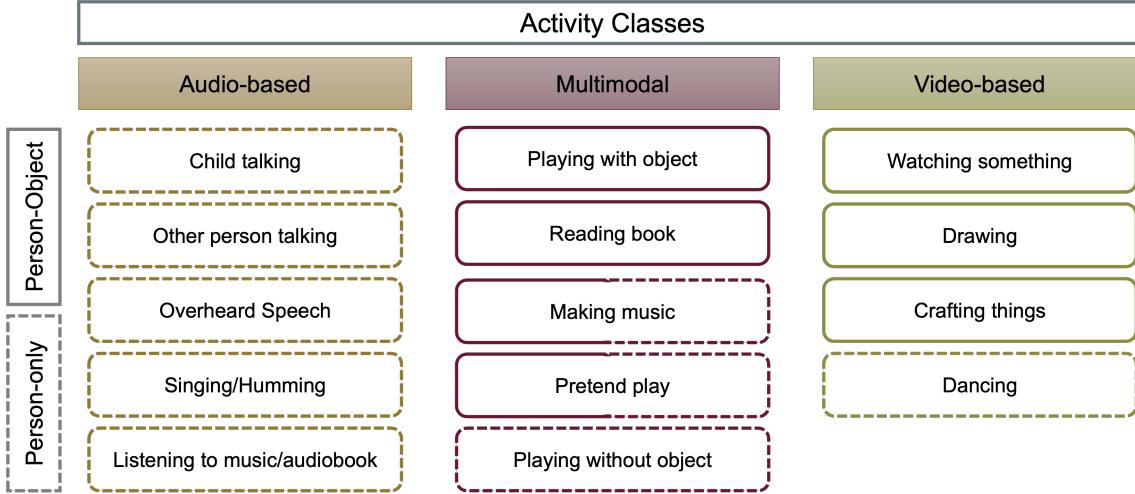
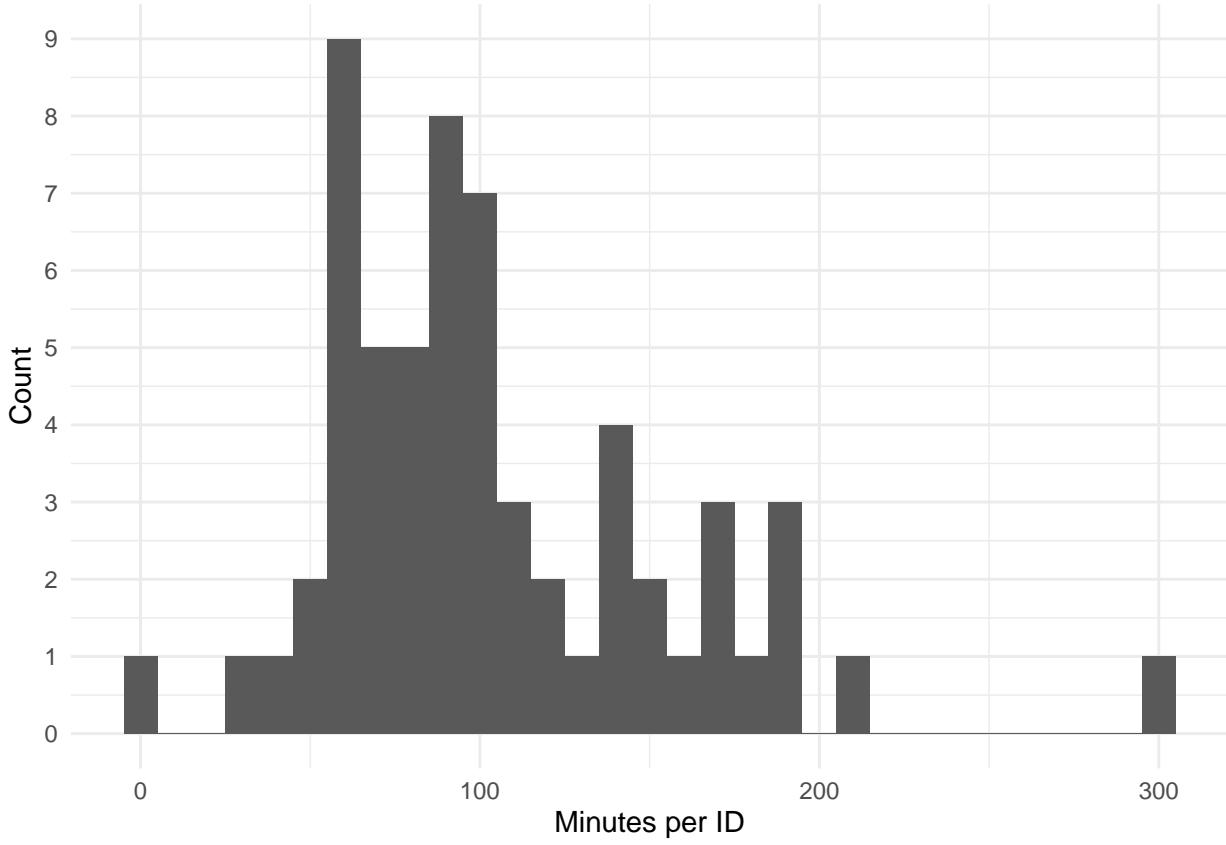


Figure 1. ChildLens Activity Class Categories

duration per child can be found in figure 2. This diverse dataset also includes a varying number of clips for each of the 14 activity classes, ranging from **x** to **x** clips per class. The clip duration depends on the activity; for example, audio-related actions like “child talking” may only last a few seconds, while activities like “reading a book” may last several minutes. As shown in table ??, the total number of **xx** clips is divided into **xx** training clips, **xx** validation clips, and **xx** testing clips for each class.

***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.



*Figure 2.* Video recording duration (in minutes) per child ID.

## Dataset Generation

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 61 children from families living in a mid-sized city in Germany, participated in the study. The videos were captured at home using a camera embedded in a vest worn by the children, which can be seen in figure 3. This setup allowed the children to move freely throughout their homes while recording their activities. The camera, a

*PatrolEyes WiFi HD Infrared Police Body Camera*, was equipped with a 140-degree wide-angle lens and captured everything within the child's field of view with a resolution of 1920x1080p at 30 fps. The camera also recorded audio, allowing us to capture the child's speech and other sounds in the environment. 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 capturing everyday activities that children typically engage in. Parents were therefore asked to include the following activities in the recordings:

- 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

The videos were recorded over a period of 12 months, resulting in a total of 106.10 hours of video footage.



Figure 3. Vest with the embedded camera worn by the children

## Step 2: Creation of Labeling Strategy

To create a comprehensive labeling strategy for the ChildLens dataset, we first defined a list of activities that children typically engage in. This list was based on previous research on child development and the activities that children are known to participate in. We then developed a detailed catalog of activities that were likely to be captured in the videos and chose to make the activity classes more granular by distinguishing between activities like “making music” and “singing/humming” or “drawing” and “crafting things”.

After an initial review of the videos, we decided to add another class “overheard speech” to capture situations in which the child is not directly involved in a conversation but can hear it. We also added “pretend play” as a separate class to capture situations in which 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.



*Figure 4.* SuperAnnotate platform utilized for video annotation

### Step 3: Manual Labeling Process

Before the actual annotation process, a setup meeting was held to introduce the annotators to the labeling strategy. To familiarize themselves with the task, the annotators were assigned 25 sample videos to practice and gain hands-on experience. These initial annotations were reviewed by us, and feedback was provided to refine their approach. A total of three feedback loops were conducted to ensure that the annotators follow the labeling strategy properly.

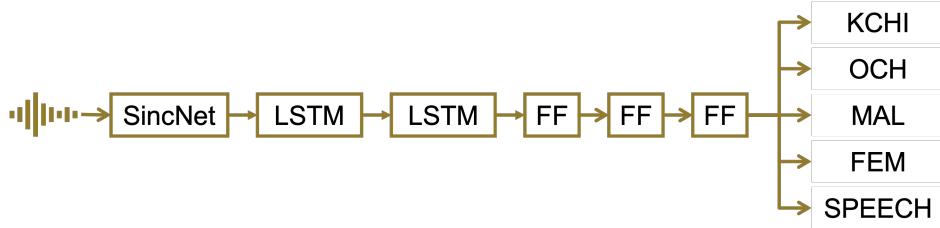
The videos were manually annotated by native German speakers who watched each video and labeled the activities present in the footage. The annotators marked the start and end points of each activity, ensuring that the annotations were accurate and detailed. The labeling process was conducted using the SuperAnnotate platform, which allowed for efficient annotation and review of the videos. Figure 4 provides a screenshot of the SuperAnnotate platform used for video annotation. To ensure the quality of the annotations the following steps were taken:

1. **Initial round of annotations:** Each set of videos is assigned to specific annotators, who handle the annotations, make changes, and apply corrections as needed. In total, three annotators were actively working on the annotation process.
2. **Quality assurance:** One person is dedicated to quality assurance, ensuring that the annotations are accurate and consistent across all videos.
3. **Review process:** After the initial annotations are completed, the annotations are reviewed by the internal team to ensure that they are accurate and complete. Any discrepancies or errors are corrected before the final submission.

### Benchmark Performance

In this chapter we present the results of two model architectures applied to the ChildLens dataset. While the dataset can be used for multimodal activity analysis, we

focus on two specific tasks: temporal activity localization using video data and voice type classification using audio data. We apply the Boundary-Matching Network (BMN) for temporal activity localization and the Voice Type Classifier (VTC) for voice type classification. Both models were trained from scratch on the ChildLens dataset to provide initial results and a benchmark for future research. A short architecture overview of the models is presented in figure ?? and figure 6.



*Figure 5.* Voice Type Classifier architecture applied to ChildLens audio data

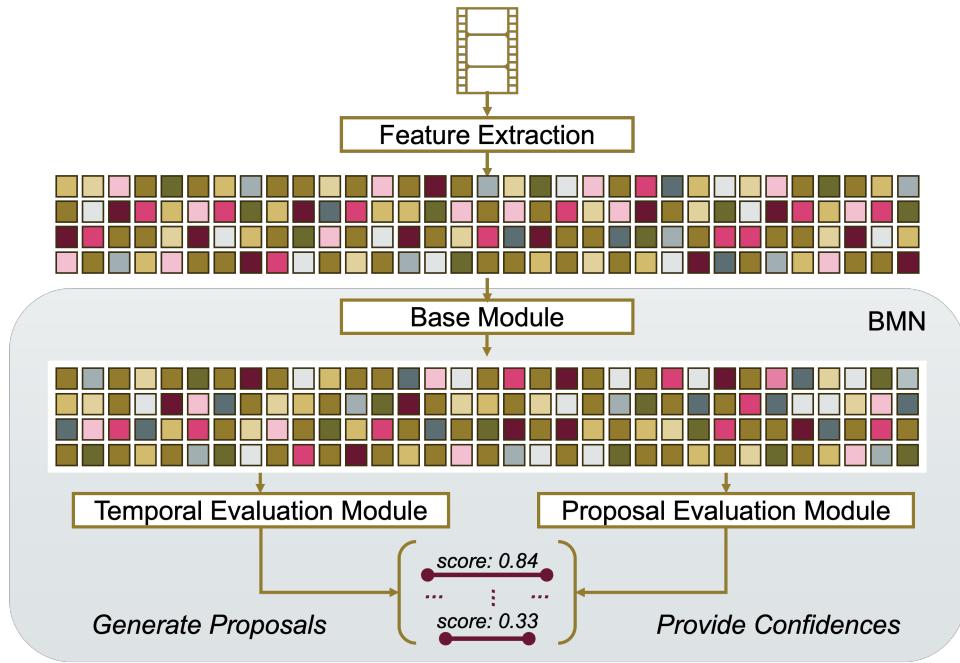
## Implementation details

### 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).



*Figure 6.* Boundary Matching Network architecture applied to ChildLens video data

## Conclusion

## Discussion

### Dataset bias

### General Discussion

## References

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>

## 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:** The child is playing with an object, such as a toy or a ball. (x clips)
2. **playing without object:** The child is playing without an object, such as playing hide and seek or catch. (x clips)
3. **pretend play:** The child is engaged in imaginative play, such as pretending to be a doctor or a firefighter. (x clips)
4. **watching something:** The child is watching a movie, TV show, or video on either a screen or a device. (x clips)
5. **reading book:** The child is reading a book or looking at pictures in a book (x clips)
6. **child talking:** The child is talking to themselves or to someone else (x clips)
7. **other person talking:** Another person is talking to the child. (x clips)
8. **overheard speech:** Conversations that the child can hear but is not directly involved in. (x clips)
9. **drawing:** The child is drawing or coloring a picture. (x clips)
10. **crafting things:** The child is engaged in a craft activity, such as making a bracelet or decoration. (x clips)
11. **singing / humming:** The child is singing or humming a song or a melody. (x clips)
12. **making music:** The child is playing a musical instrument or making music in another way (x clips)
13. **dancing:** The child is dancing to music or moving to a rhythm. (x clips)
14. **listening to music / audiobook:** The child is listening to music or an audiobook. (x clips)

**List of ChildLens Location Classes**

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