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

- TODO: research about egocentric video datasets

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 a brief description of each class 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*.

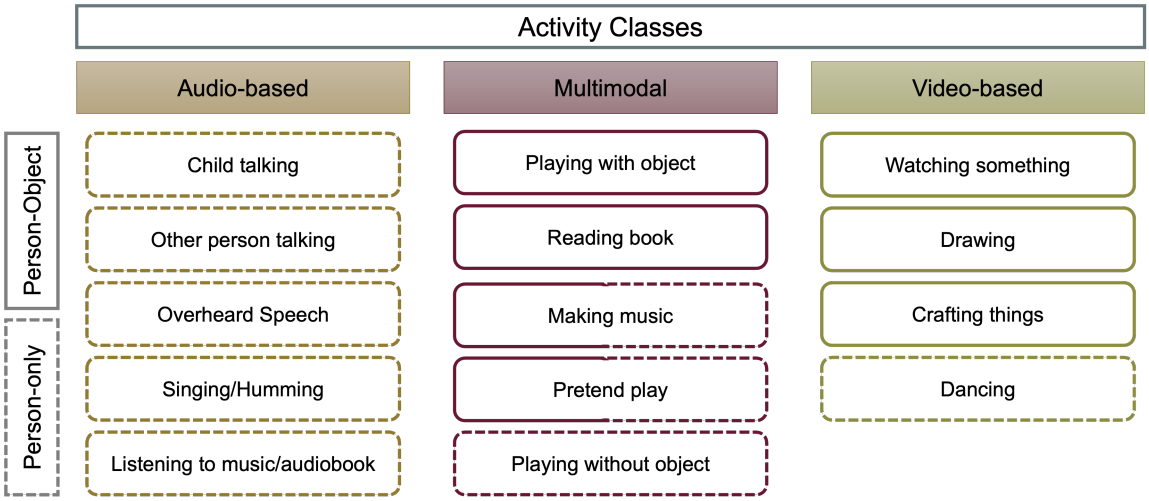

Figure 1. ChildLens Activity Class Categories

Statistics. The ChildLens dataset comprises of 343 video files with a total of 106.10 hours recorded by 61 children aged 3 to 5 years (M=4.52, SD=0.92). It includes 107 videos from children aged 3, 122 videos from children aged 4, and 114 videos from children aged 5. The duration of recorded video material per child varies between 4.03 and 303.42 minutes (M=104.37, SD=51.65). A detailed distribution of the video duration per child can be found in figure 3. This diverse dataset also includes a varying number of clips [**MB**: unklar was clips sind - das kommt ja später noch oft vor, von dem her gerne definieren.] for each of the 14 activity classes, ranging from **x** to **x** clips per class [**MB**: nicht klar was du hier meinst. sind das instances für jede class oder unique annotations?]. 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

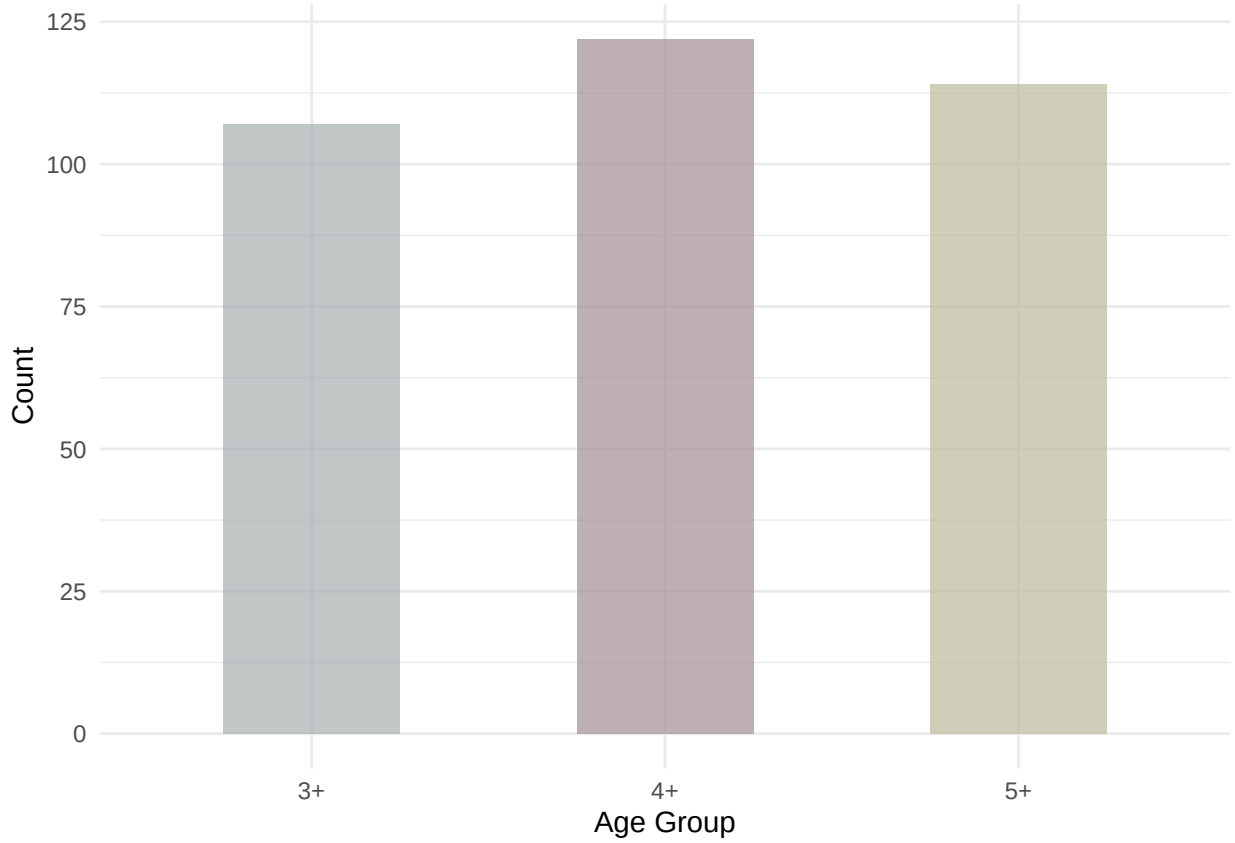

Figure 2. Number of videos per age group.

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.

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.

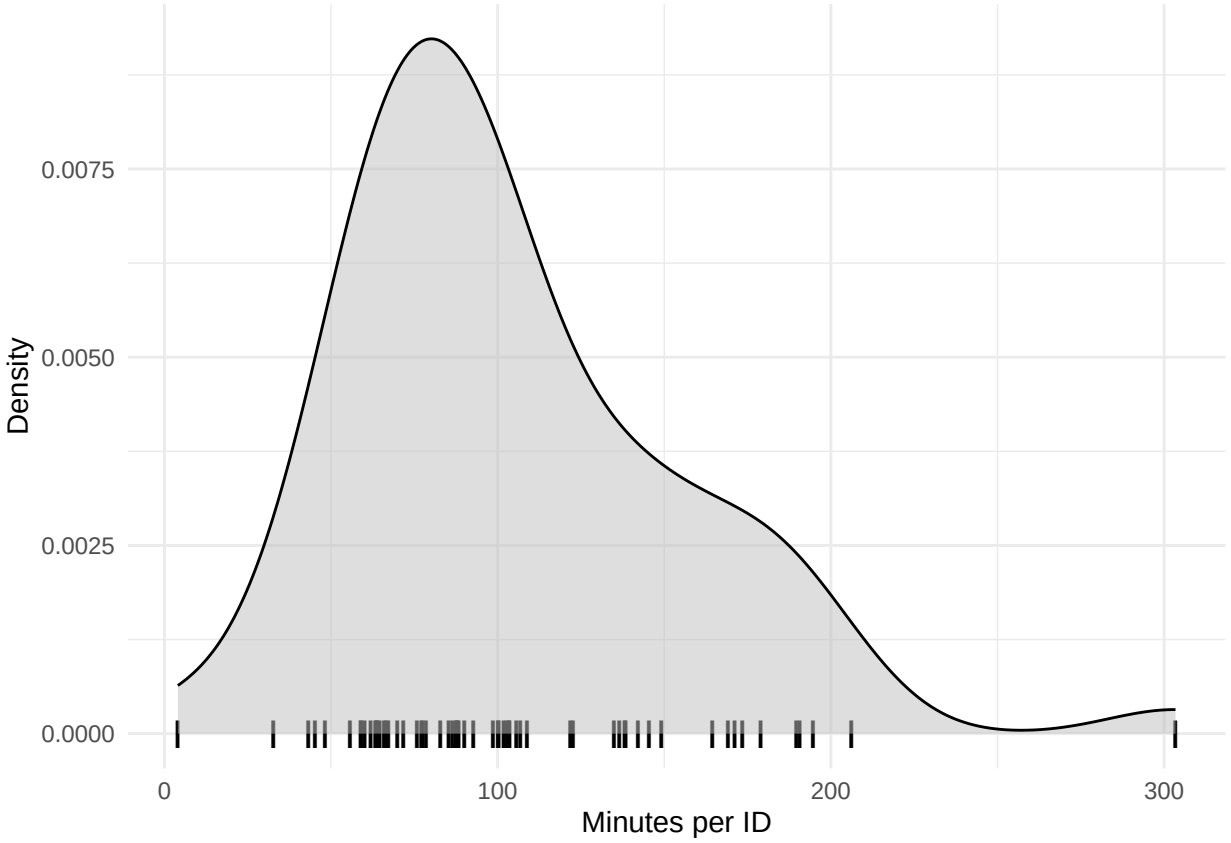

Figure 3. Video recording duration (in minutes) per child ID.

Step 1: Collection of Egocentric Videos

The ChildLens dataset consists of egocentric videos recorded by children aged 3 to 5 years over a period of 12 months. 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 4. 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 elements in the recordings:

- Child spends time in different rooms and performs various activities in each room
- 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

Figure 4. 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 5. 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 the research team, and feedback was provided to refine the 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 5 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 applying two model architectures to the ChildLens dataset. While the dataset supports multimodal activity analysis, we focus on two specific tasks: temporal activity localization using video data and voice type classification using audio data. For temporal activity localization, we use the Boundary-Matching Network (BMN) model, a state-of-the-art approach in this domain, and train it from scratch on the unique activity classes in the ChildLens video data. For voice type classification, we apply the Voice Type Classifier (VTC) (Lavechin, Bousbib, Bredin, Dupoux, & Cristia, 2020), also state-of-the-art, which was trained on similar data. Both models provide initial results and establish a benchmark for future research.

Boundary-Matching Network

We utilize the Boundary-Matching Network model (Lin, Liu, Li, Ding, & Wen, 2019) for temporal activity localization.

Implementation Details.

- details on training
- optimization strategy
- video prepocessing etc

Evaluation. The most relevant findings of the BMN evaluation are summarized in the following sections.

- Class accuracy differences
  - full list of classification accuracy can be bound in figure **xx**
  - classes **xx** and **xx** are hardest to classify, because of **xx**
  - **xx** classes were easier to classify than others such as **xx** and **xx** because of **xx**
- Class confusion
  - describe which classes are often confused with each other (e.g. **xx** and **xx**)

Voice Type Classifier

The Voice Type Classifier (Lavechin et al., 2020) is a state-of-the-art model designed to classify audio rawfiles into five distinct voice types: *key child*, *other child*, *male speech*, *female speech*, and *speech.* Its architecture processes audio by first dividing it into 2-second chunks, which are passed through a SincNet to extract low-level features. These features are then fed into a stack of two bi-directional LSTMs, followed by three feed-forward layers. The output layer uses a sigmoid activation function to produce a score between 0

and 1 for each class. The VTC is trained on 260 hours of audio material obtained from different child-centered audio datasets. Model valuation is performed by utilizing the F_1-measure, which combines precision and recall using the following formula:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

where precision $= \frac{\text{tp}}{\text{tp+fp}}$ and recall $= \frac{\text{tp}}{\text{tp+fn}}$ with

- tp being the number of true positives,
- fp being the number of false positives, and
- fn being the number of false negatives.

The F_1 is a metric that combines precision and recall into a single value, calculated as their harmonic mean. It ranges from 0 to 1, with 1 representing perfect precision and recall, and 0 indicating no correct prediction The interpretation of the F_1 score depends on the specific application of the model. Generally, an F_1 score above 0.8 is considered good, while values above 0.9 are considered excellent. In some cases, a score around 0.5 can still be deemed acceptable, depending on the balance between precision and recall.. The F_1 score is computed for each class and averaged to provide an overall measure. No collar is applied to the evaluation, meaning that the prediction have to be exact to be considered correct. The model achieves an F_1 score of 57.3, outperforming the previous state-of-the-art LENA model by 10.6 points.

Evaluation. The most relevant findings of the VTC evaluation are summarized in the following sections. - summarize most crucial findings

- which class performs worst - maybe why?
- which is best?
- how is the average compared to the ACLEW-Random dataset

Table 1

Comparison of VTC performance on the ACLEW-Random dataset (used for model evaluation) and the ChildLens dataset, highlighting the F1 measure for each class and the average F1 score

Dataset	KCHI	OCH	MAL	FEM	SPEECH	AVG
ACLEW-Random	68.7	33.2	42.9	63.4	78.4	57.3
ChildLens	59.1	79.2	17.8	33.4	68.3	51.5

Discussion

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 Bias

Overall, the dataset demographics are balanced. From the 61 children who participated in the study, 32 children are female and 29 are male.

- is there gender bias in the dataset itself (how many female how many male)
- is ther gender bias in some categories (e.g. more female for drawing etc.)

General Discussion

Conclusion

In this paper, we introduced the ChildLens dataset, a novel egocentric video dataset designed for activity analysis in children. The dataset contains a wide range of children's daily live activities, captured in naturalistic environments. We outlined the data collection

process and the generation of activity labels, providing detailed information on the labeling strategy employed. Initial results of applying two state-of-the-art models to the dataset were presented, establishing a benchmark for future research. While our current analysis treats audio and video independently, future studies could leverage multimodal approaches to gain deeper insights into children's behavior and activity patterns, advancing the understanding of developmental and interactional contexts.

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 mekody. (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