Exploring Aspects of Social Interaction using Machine Learning

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

tbd

Exploring Aspects of Social Interaction using Machine Learning

Introduction

Methodology

The Quantex dataset includes

Dataset Description

Statistics. The Quantex dataset contains a total of 197.20 hours of video footage from 503 video recordings, collected by 76 children aged 3 to 5 years (M=4.53, SD=0.81). The children were grouped into three age categories, with 167 videos being recorded of children age 3, 180 videos for children age 4, and 156 videos at age 5. Individual recording durations vary widely, ranging from 10.43 to 391.18 minutes per child (M=155.68, SD=82.62). Figure 1 illustrates the detailed distribution of recording lengths, reflecting the diversity in individual contributions to the dataset.

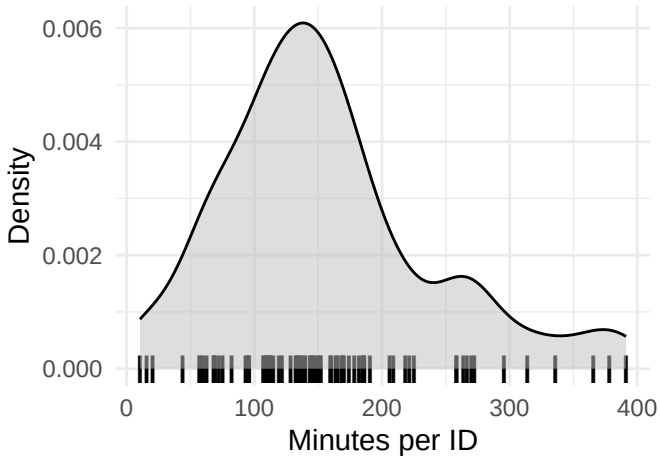

Figure 1. Video recording duration (in minutes) per Child in the Quantex Dataset.

Annotation Strategy. The dataset annotations cover four key elements: persons, faces, objects the child interacts with, and gaze direction. Gaze information identifies

whether a detected person's gaze is directed toward the child or not. For every detected person (or reflection of a person, such as in a mirror) and face, additional attributes like age and gender are collected. Objects are categorized into six distinct groups: book, screen, animal, food, toy, and kitchenware, with an additional category for other objects. The dataset focus is on detecting and labeling instances of (social) interaction and engagement through these key categories. The annotation strategy is displayed in Figure 2.

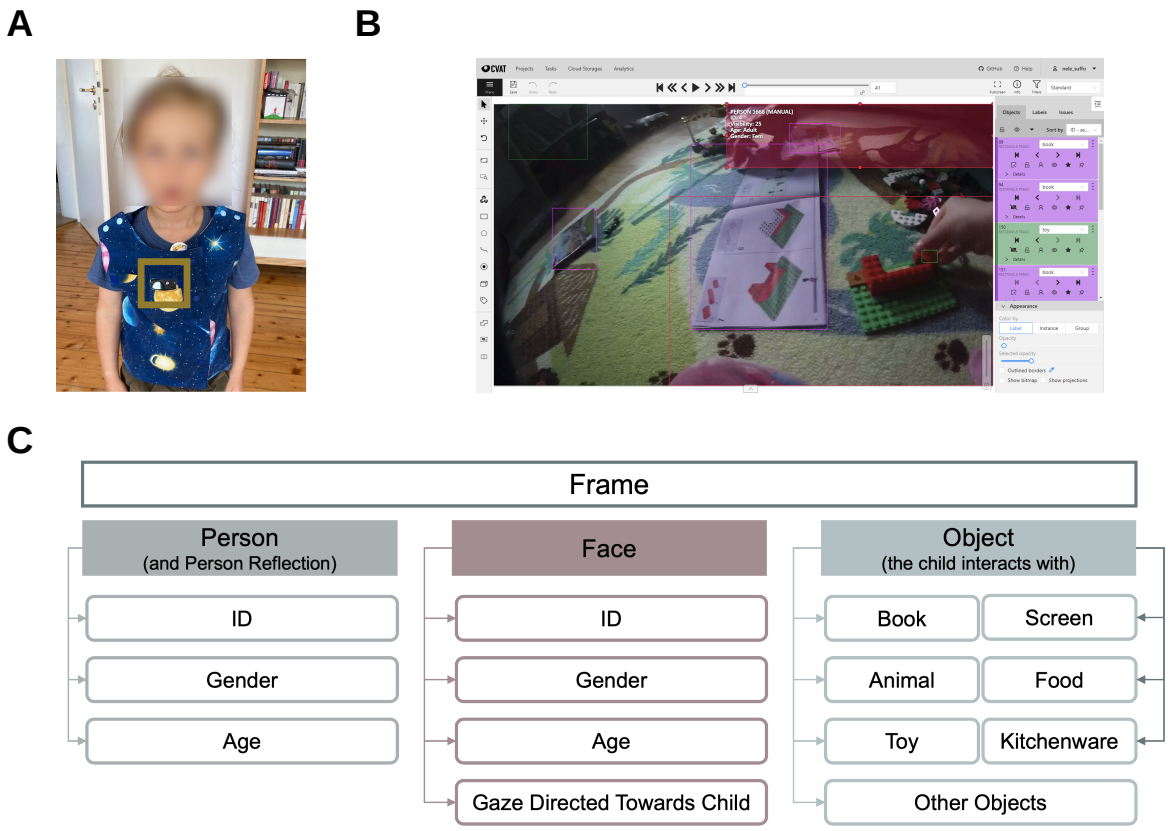

Figure 2. **A** – Vest with the embedded camera worn by the children, **B** – CVAT platform utilized for video annotation, **C** – Annotation strategy in the Quantex dataset.

Data Collection

This study collected egocentric video recordings from 76 children, aged 3 to 5 years, over a span of 73 months. Participating families lived in a mid-sized city in Germany. To

capture the children's everyday experiences, a wearable vest equipped with a camera was used, as shown in figure 2. The camera, a *PatrolEyes WiFi HD Infrared Police Body Camera*, provided high-definition video (1920x1080p at 30 fps) with a 140-degree wide-angle lens and also recorded audio. Children were free to move around and engage in their usual activities at home without any interference or instructions given to their parents.

Data Preprocessing

For the video data, the annotation strategy required persons, faces, and objects to be labeled even when only partially visible, as long as key features such as facial landmarks (e.g., nose, eye, or mouth) or parts of a person or object were clearly visible. Frames that were too blurry due to movement were marked as "noise" and excluded from further analysis. Additionally, frames where the child was not wearing the camera, as well as any scenes containing nudity, were also labeled as noise and removed from the dataset. To prepare the video data for analysis, one frame per second was annotated, corresponding to every 30th frame in the video. Similarly, every 30th raw frame was extracted from the annotated video files. No preprocessing was applied to the audio data, which was used in its raw form for analysis.

Automated Analysis Pipeline

Person Detection.

Face Detection.

Gaze Classification.

Voice Detection and Classification.

Feature Extraction

Results

Presence of Aspects of Social Interaction

Presence of a Person.

Presence of a Face.

Presence of Gaze Directed at the Child.

Presence of Language.

Co-occurrence of Aspects of Social Interaction

General Discussion

We present the ChildLens dataset, a unique egocentric video-audio dataset that documents children's everyday experiences, spanning a wide age range of 3 to 5 years. This dataset is particularly distinctive due to its diversity in terms of the number of children it includes and the variety of activity labels it covers. By focusing on both visual and auditory data, the ChildLens dataset provides comprehensive annotations for a broad spectrum of key activities, offering rich insights into children's social and behavioral interactions in naturalistic settings. This makes it a valuable resource for studying developmental processes in children, with a focus on their cognitive, emotional, and social growth.

In comparison to other freely available datasets, the ChildLens dataset stands out due to its broad age span and diverse set of activity labels. Most other datasets focus either on toddlers, are limited to dyadic interactions or were recorded in lab settings, with all of them lacking a comprehensive range of activity labels. Furthermore, most of these datasets either capture only audio or video, missing the multimodal aspect crucial for understanding children's everyday experiences. In contrast, ChildLens includes naturalistic

recordings from children's home environments, over an extended period, and features a variety of activity types. The dataset also captures whether children are engaged in activities alone or with others and provides detailed demographic information about all individuals involved. This comprehensive approach enables a deeper and more holistic understanding of children's interactions and developmental trajectories.

The usefulness of the ChildLens dataset is demonstrated by its successful application to well-established models. For example, the pretrained Voice-Type Classifier for audio transcription achieves performance comparable to previous datasets, while the Boundary-Matching Network (BMN) produces robust results for activity localization, consistent with its performance on other datasets. These results indicate that the ChildLens dataset's annotations align well with model predictions, highlighting its quality and potential for multimodal research. Moreover, the successful application of these models demonstrates how the dataset can support and automate the analysis of children's everyday activities.

Expanding the potential for multi-method approaches, activity localization could be further enhanced by incorporating object identification, allowing for better tracking of the objects children interact with during daily routines, as demonstrated in adult-focused studies (Kazakos, Huh, Nagrani, Zisserman, & Damen, 2021). Research by Bambach et al. (Bambach, Lee, Crandall, & Yu, 2015) also emphasizes the importance of hand detection in egocentric video for activity recognition. Their work on using Convolutional Neural Networks for hand segmentation highlights how such techniques can differentiate activities, offering a deeper understanding of children's interactions and behaviors.

The integration of visual and auditory data in the ChildLens dataset enables a more detailed and comprehensive understanding of children's daily experiences. Complex activities such as pretend play and reading a book, which require both audio and video for accurate detection, exemplify the strength of this multimodal approach. While previous

studies, such as those analyzing disfluencies in children's speech during computer game play (Yildirim & Narayanan, 2009), have demonstrated that combining visual and auditory information can improve performance, few studies have explored this in the context of children's naturalistic activities. With ChildLens, the combination of naturalistic data and multimodal analysis creates new opportunities for in-depth insights into children's cognitive, emotional, and social development, particularly for activities best captured through both modalities.

Despite its strengths, one limitation of the ChildLens dataset is the class imbalance, especially in underrepresented activity classes, which could affect model training and evaluation. More frequent activities, such as "child talking" (7447 instances, 649 minutes) and "playing with object" (317 instances, 1371 minutes), dominate the dataset, while less common activities like "dancing" (2 instances, 0.57 minutes) and "making music" (2 instances, 2.13 minutes) are scarcely represented. Similarly, activities like "pretend play" (59 instances, 158.84 minutes) and "reading a book" (81 instances, 328.70 minutes) appear less frequently. This imbalance may lead to skewed model performance, making it harder to accurately classify rare activities. Possible solutions to this challenge could involve merging rare activity classes into broader categories or excluding them from model training, though these approaches may reduce the dataset's diversity. Other methods, such as resampling or augmentation, could help balance the dataset and improve model performance (Alani, Cosma, & Taherkhani, 2020; Spelmen & Porkodi, 2018).

In addition to class imbalance, another potential limitation is the sampling bias. Since the recordings are largely influenced by parental decisions about when and how often activities are captured, certain activities or settings may be overrepresented or underrepresented based on these preferences. Furthermore, the dataset primarily focuses on families from a mid-sized German city, limiting its geographic and cultural diversity. Expanding the dataset to include a broader range of families from different regions and cultures would enhance its generalizability and applicability to various research contexts.

The study of children's everyday experiences is crucial for understanding their cognitive, emotional, and social development. These daily interactions provide important insights into how children learn, grow, and engage with their environment. The ChildLens dataset makes a valuable contribution to this field by offering a rich multimodal resource that captures the complexities of children's lives in naturalistic settings. With its comprehensive annotations and potential to automate the analysis of children's activities, the dataset enables researchers to gain deeper and more detailed insights into children's development. By making such analyses more efficient and accessible, the ChildLens dataset creates new opportunities for understanding the complexities of early childhood development and provides a foundation for future research in this area.

References

Alani, A. A., Cosma, G., & Taherkhani, A. (2020). Classifying Imbalanced Multi-modal Sensor Data for Human Activity Recognition in a Smart Home using Deep Learning. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–8. Glasgow, United Kingdom: IEEE. https://doi.org/10.1109/IJCNN48605.2020.9207697

Bambach, S., Lee, S., Crandall, D. J., & Yu, C. (2015). Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1949–1957. Santiago, Chile: IEEE. https://doi.org/10.1109/ICCV.2015.226

Kazakos, E., Huh, J., Nagrani, A., Zisserman, A., & Damen, D. (2021). With a Little Help from my Temporal Context: Multimodal Egocentric Action Recognition. https://doi.org/10.48550/ARXIV.2111.01024

Spelmen, V. S., & Porkodi, R. (2018). A Review on Handling Imbalanced Data. *2018 International Conference on Current Trends Towards Converging Technologies (ICCTCT)*, 1–11. Coimbatore: IEEE. https://doi.org/10.1109/ICCTCT.2018.8551020

Yildirim, S., & Narayanan, S. (2009). Automatic Detection of Disfluency Boundaries in Spontaneous Speech of Children Using Audio–Visual Information. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(1), 2–12. https://doi.org/10.1109/TASL.2008.2006728

Appendix

List of ChildLens Activity Classes

The dataset contains the following list of activities.

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

List of ChildLens Location Classes

1. livingroom

Table 1

Number of video instances and the total duration (in minutes).

Category	Activity Class	Instance Count	Total Duration (min)
Audio	Child talking	7447	649.10
	Other person talking	6113	455.29
	Overheard Speech	1898	299.44
	Singing/Humming	277	82.00
	Listening to music/audiobook	68	222.14
Video	Watching something	2	5.09
	Drawing	62	374.91
	Crafting things	26	109.14
	Dancing	2	0.57
Multimodal	Playing with object	317	1371.06
	Playing without object	25	28.87
	Pretend play	59	158.84
	Reading a book	81	328.70
	Making music	3	2.13

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

Activity Class Statistics