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.

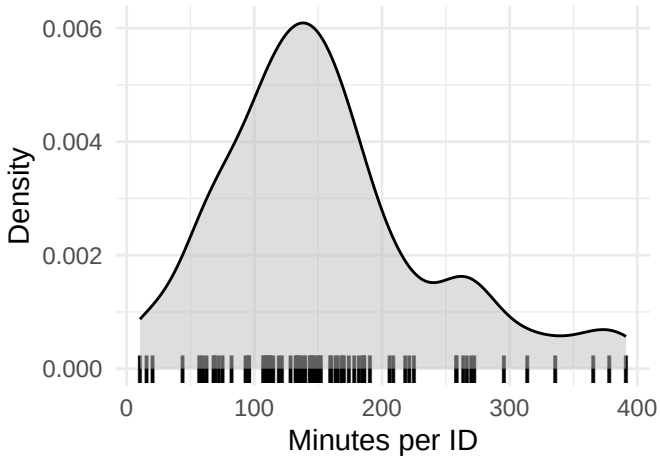

Figure 1. Video recording duration (in minutes) per Child in the Quantex Dataset.

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.

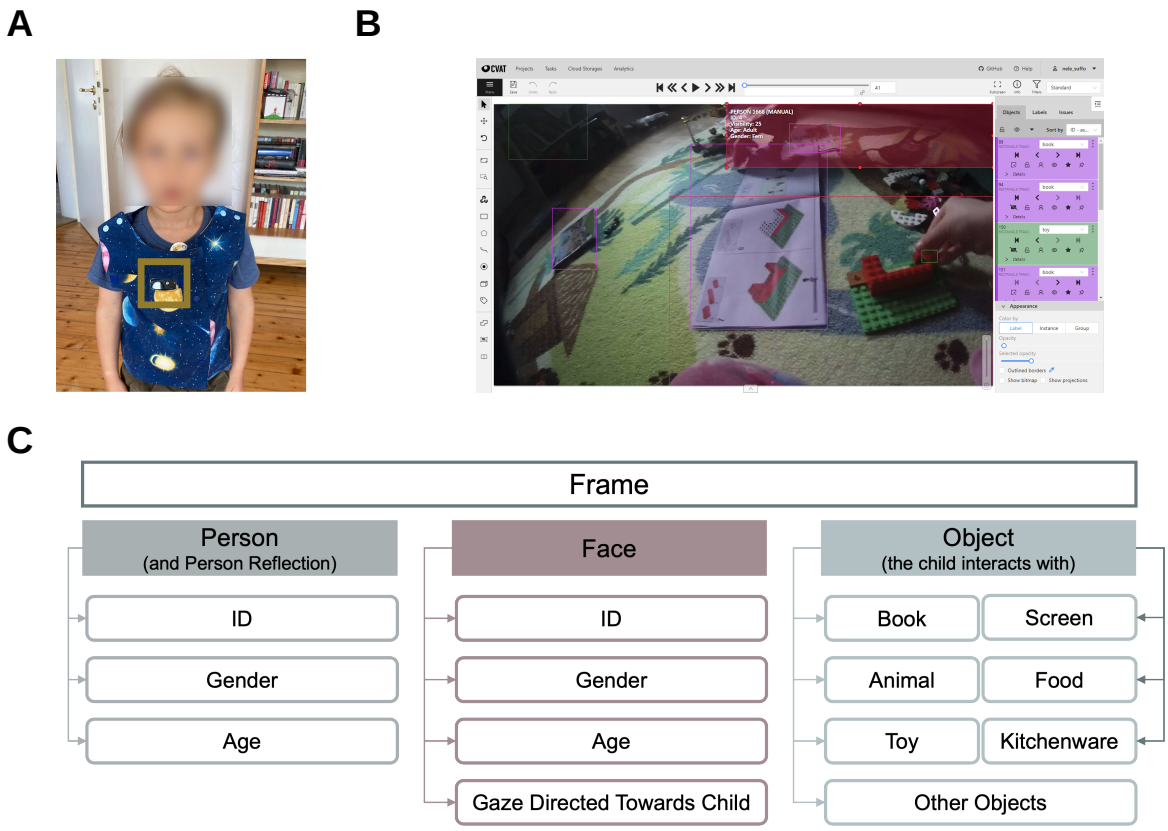

Figure 2. **A** – Vest with the embedded camera worn by the children, **B** – CVAT platform utilized for video annotation, **C** – Annotation strategy in the Quantex dataset.

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. We employed a YOLOv11 model pretrained for face detection (Codd, 2024), which was fine-tuned on our dataset to adapt it to the unique characteristics of our egocentric dataset, captured using chest-mounted cameras. While we initially experimented with the MTCNN model, its performance on our dataset proved insufficient. Consequently, we chose YOLO due to its streamlined training process and fewer

Table 1

Evaluation metrics for the Yolov11 face detection model trained on the Quantex dataset.

Dataset	Precision	Recall	F1-Score	Accuracy	False Positive Rate	False Negative Rate
ChildLens	77.43	69.21	73.07	85.00	0.03	0.26

requirements for data preparation. The dataset was divided into 80% for training, 10% for validation, and 10% for testing, consisting of 64 videos in total. This split corresponded to 72,687 frames for training, 7,720 frames for validation, and 9,272 frames for testing.

Model training was conducted using the Ultralytics framework (Jocher, Jing, & Chaurasia, 2023) on a Linux server equipped with 48 cores and 187 GB of RAM. The training process utilized YOLO's built-in data augmentation, a batch size of 16, a cosine annealing learning rate scheduler, and early stopping after 10 epochs without improvement, with a maximum of 200 epochs. Training concluded after 86 epochs, achieving a mean average precision (mAP) of 0.89 on the validation set. The most relevant model evaluation metrics are summarized in Table 1.

The model is an essential part of our automated analysis pipeline,detecting faces in each frame for subsequent gaze analysis. An examination of false negatives revealed that the model struggled in scenarios involving occlusions, such as rapid movements by the child, camera shake, parts of the child obscuring another person's face, or the presence of small faces in the background. However, these instances are not central to our study, as they are unlikely to involve individuals actively interacting with the child.

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

References

Codd, A. (2024). *YOLOv11n-face-detection*. Retrieved from https://huggingface.co/AdamCodd/YOLOv11n-face-detection

Jocher, G., Jing, Q., & Chaurasia, A. (2023). *Ultralytics YOLO*. Retrieved from https://github.com/ultralytics/ultralytics

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 2

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