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. From the 503 video files, 100 videos were annotated manually as training input for the automated analysis pipeline.

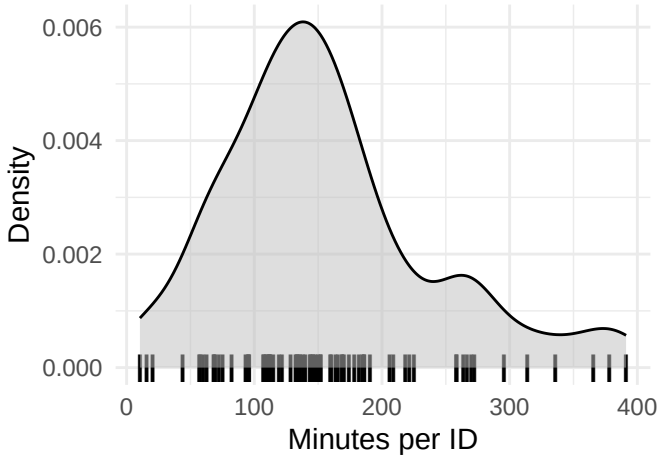

Figure 1. Video recording duration (in minutes) per Child in the Quantex Dataset.

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

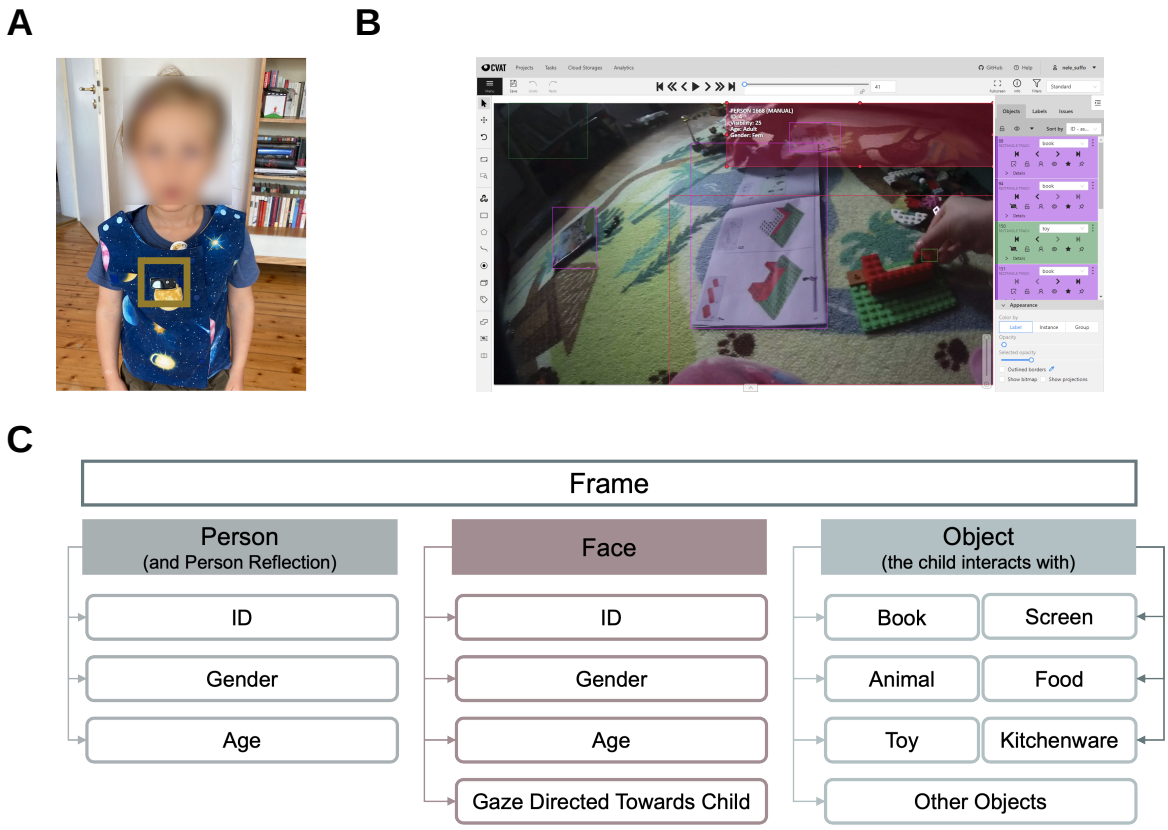

Figure 2. **A** – Vest with the embedded camera worn by the children, **B** – CVAT platform utilized for video annotation, **C** – Annotation Strategy in the Quantex dataset.

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 labele d 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 requirements for data preparation. The 100 annotated videos were divided into 70% for training, 10% for validation, and 20% for testing. This split corresponded to 51 with 72687 frames for training, 6 videos with 7720 frames for validation, and 7 videos with 9272 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 precision of 0.90 and a recall of 0.83, resulting in an F_1-score of 0.86 on the testing set. This indicates strong performance in correctly identifying most faces while minimizing errors, although some challenges remain. These performance metrics, summarized in Table 1, underscore the model's ability to reliably detect faces, with further details and evaluation available in the supplementary materials.

The model performed well in detecting faces, particularly when fully visible from the front, but few challenges remain in more dynamic scenarios. For example, faces that are partially visible, rotated, or seen from the side often resulted in detection errors. Furthermore, false negatives were more common when faces were occluded by the child's body, blurred due to movement, or situated in the background. While background faces are less relevant, as they are unlikely to be part of an interaction with the child, missed detections due to occlusions or motion blur present a greater challenge. In these cases, we rely on adjacent frames to provide clearer views for more accurate classification. These difficulties underscore the challenges of working with egocentric video data, where dynamic movement and varying perspectives, typical of chest-mounted camera recordings, introduce

Table 1

Evaluation metrics for the YOLOv11 face detection model trained on the Quantex dataset.

Dataset	Precision	Recall	F1-Score
Quantex	0.90	0.83	0.86

additional complexity.

Accurate face detection remains a crucial step in our automated analysis pipeline, as it serves as the foundation for subsequent gaze classification. Identifying the presence and position of faces ensures that gaze direction can be reliably analyzed, allowing us to determine when and how individuals engage 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

Supplementary Material

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