

Exploring Aspects of Social Interaction using Machine Learning

Nele-Pauline Suffo¹, Pierre-Etienne Martin², Anam Zahra², Daniel Haun², & 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

Childrens everyday experiences are known to shape childrens development but only few studies investigate how children actually spent their time at home in naturalistic setting. More particular we were interested in how children's social interactions with others or interactions with objects are observable in their everyday life. To do so, we utilized the Quantex Dataset, an egocentric video and audio dataset of children aged 3-5 years, to investigate the presence of persons, faces, gaze, and objects in children's everyday interactions. We trained a YOLO11 model to detect persons and faces in the videos and analyzed the presence of gaze and objects in the videos. We furthermore applied a pre-trained voice type classifier to detect speech in the audio data. Our results show that children's everyday interactions are characterized by the presence of persons and faces, with the child's gaze directed towards others in 60% of the interactions. Additionally, children interacted with objects in 40% of the videos, with toys being the most common object category. Agr group analysis revealed that children aged 3 years showed more interactions with objects compared to older children. Our findings provide insights into the diversity of children's everyday experiences and highlight the importance of multimodal data for understanding children's social interactions and engagement.

Keywords: Quantex Dataset, egocentric video, audio dataset, children, social interactions, object interactions, gaze, multimodal data, computer vision, audio analysis, developmental psychology

Exploring Aspects of Social Interaction using Machine Learning

Introduction

According to various developmental psychologists, children's everyday experiences play a vital role in their development (Carpendale & Lewis, 2020; Heyes, 2018; Piaget, 1964; Rogoff, Dahl, & Callanan, 2018; Smith, Jayaraman, Clerkin, & Yu, 2018; Tomasello, 2009; Vygotsky, 1978). Everyday interactions, in particular, have been recognized for decades as crucial in the process of actively constructing knowledge (Piaget, 1964) and in transforming sensory experiences into structured understanding (Vygotsky, 1978). Building upon these foundational theories, more recent research has examined the mechanisms of social interaction further. For instance, Tomasello (2009) introduced the concept of shared intentionality, illustrating how collaborative activities enable children to comprehend others' intentions and perspectives, leading to cooperative behaviors and cultural learning .

Whereas theoretical frameworks and controlled laboratory studies have significantly advanced our understanding of children's social development, they often fail to capture the complexities of interactions occurring in naturalistic settings. Observing children in their everyday environments offers a more authentic view of their social behaviors; however, this approach presents challenges due to the extensive data collection and analysis required.

To address these challenges, researchers have increasingly turned to data-driven approaches that utilize sensors and recording devices to gather objective data on social interactions. For instance, Onnela, Waber, Pentland, Schnorf, and Lazer (2014) employed wearable sensors to analyze social interactions in adult work settings, capturing the duration of close proximity between individuals. The study inferred that women were more talkative than men and more likely to be physically close to other women in group settings. Rossano et al. (2022) examined social interactions among 31 two- to four-year-olds using 563 hours of video and audio recordings from a preschool during free play sessions over seven days. Manual interaction labels revealed that four-year-olds engaged in more

cooperative social interactions and experienced fewer conflicts than two-year-olds, with object play and conversations being the most common forms of social engagement in both age groups. Dai et al. (2022) investigated social interactions of 174 preschool children over three years, collecting voice and proximity data using wearable wireless RFID tags to study the co-development of social interactions and language acquisition. They employed manually labeled interaction data to train a temporal segment model that automatically identified periods of free play or class play, concluding that classmates frequently engaged in both contexts. Lemaignan, Edmunds, Senft, and Belpaeme (2018) created a dataset comprising 45 hours of manually labeled social interactions between 45 child-child pairs and 30 child-robot pairs, including video and audio recordings, 3D facial data, skeletal information, and game interactions. By not providing specific instructions to the children, the researchers aimed to capture interactions in naturalistic settings. However, each laboratory session was limited to 40 minutes.

While these studies have advanced our understanding of social interactions, they often focus on controlled environments or are constrained by limited observation periods. Moreover, the manual data collection and analysis involved remain labor-intensive and time-consuming, and the current body of research lacks comprehensive data-driven studies analyzing children's social interactions within their home environments.

To overcome these limitations, our study investigates social interactions in naturalistic home settings over an extended period. We have created the **Quantex** dataset which currently includes 197.20 hours of egocentric video and audio recordings from children aged 3 to 5 years and enables the analysis of specific patterns of social interactions, including:

- Presence of Individuals: Utilizing Yolo11 for person detection to identify when others are present in the child's environment.
- Presence of Faces: Employing Yolo11 face detection to recognize faces the child encounters.

- Object Interactions: Analyzing the objects with which the child interacts using Yolo11 object detection
- Gaze Behaviors: Classifying gaze direction to determine when others are looking at the child.
- Speech Dynamics: Implementing voice type classification to differentiate between the child's speech and that of others, distinguishing between peers and adults.

The primary objectives of this study are to quantify interaction patterns by measuring the frequency and duration of each identified interaction type, both individually and in combination. Additionally we compare these patterns across different age groups within the 3 to 5-year range to identify developmental variations and milestones.

Understanding these interaction patterns can inform developmental psychology about the actual nature of social interactions in children's everyday lives.

Methodology

This chapter outlines the methodology used in this study to collect, annotate, and analyze video and audio recordings of children's everyday interactions. The aim of the study is to investigate key aspects of social interactions and engagement, such as the presence of persons, faces, gaze direction, and objects the child interacts with. The following sections provide a detailed description of the data collection process, the structure and characteristics of the dataset, the annotation strategy, and the preprocessing applied to the data prior to analysis. Additionally, an overview of the automated analysis pipeline is provided, giving details about the models used for person and face detection, gaze classification, object detection, and the application of a pre-trained voice type classifier.

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. Data collection is ongoing, and the number of children will continue to increase as the study progresses. The data collection process was approved by the local ethics committee, and all participating families provided written informed consent, allowing the researchers to use the data for scientific purposes. In accordance with data privacy regulations, every child was assigned a unique anonymized ID to protect their identity. Moreover, the video recordings are stored on a secure server and are only accessible to the research team, all of whom have signed confidentiality agreements.

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

As of now, the ongoing data collection process has resulted in a total of 503 video recordings, with a combined duration of 197.20 hours.

Dataset Overview

The Quantex dataset includes video and audio recordings from 76 children aged 3 to 5 years ($M=4.53$, $SD=0.81$). The dataset contains 167 videos from three-year-olds, 180 videos from four-year-olds, and 156 videos from five-year-olds. The number of videos per child varies, as parents decide when and how often to record. The recording duration per child ranges from 10.43 to 391.18 minutes ($M=155.68$, $SD=82.62$). The total duration of all video recordings in the dataset is 197.20 hours. Figure 1 shows the distribution of video duration per child.

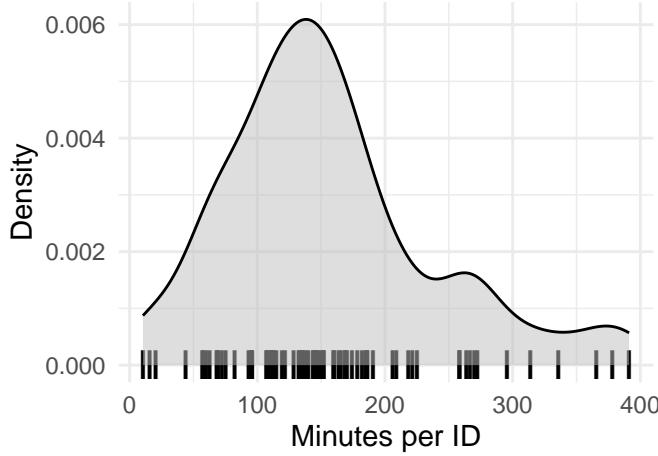


Figure 1. Video recording duration (in minutes) per Child in the Quantex Dataset.

Annotation Strategy

The dataset annotations cover four key elements: persons, faces, gaze direction, and objects the child interacts with. For each detected person (or reflection of a person, such as in a mirror) and face, additional attributes, such as age and gender, are collected. Gaze information indicates whether a detected person's gaze is directed toward the child or not. Faces are annotated even when occluded or blurry to ensure comprehensive coverage of interactions. Partially visible faces are also annotated if key facial features, such as the nose, eyes, or mouth, remain identifiable. Objects are annotated only when the child is actively interacting with them. These objects are categorized into six distinct groups: book, screen, animal, food, toy, and kitchenware, with an additional category for other objects. The annotation strategy is summarized in Figure 2.

The annotations were generated manually by a team of human annotators. Each video was randomly assigned to an initial annotator, and then reviewed by a second annotator to ensure consistency and accuracy. This peer review process helped to identify and resolve discrepancies, ensuring high-quality annotations.

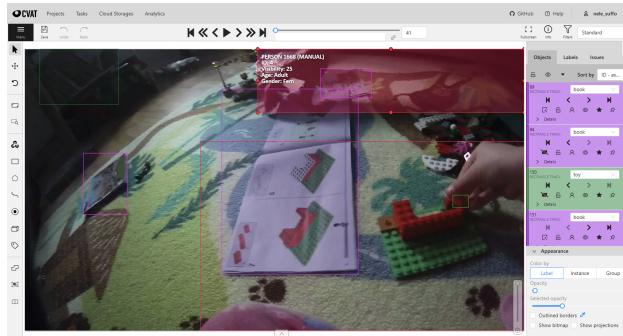
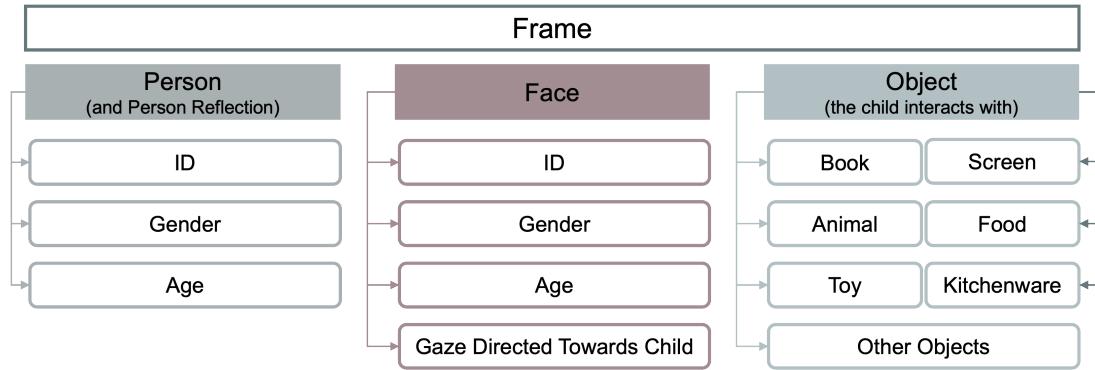
A**B****C**

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 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, eyes, or mouth) or parts of a person or object were clearly visible. To prepare the video data for analysis, one frame per second was annotated, corresponding to every 30th frame in the video. This frame sampling was chosen to balance the need for a representative sample of the video while keeping the analysis manageable. Similarly, every 30th raw frame was extracted from the annotated video files for further processing. No preprocessing was applied to the audio data, which was used in its raw form for analysis.

Automated Analysis Pipeline

Our automated analysis pipeline consists of four key modules: person and face detection, gaze classification, object detection, and voice type classification. Each module operates independently, utilizing separate machine learning models. Except for the voice type classifier, all models were trained on the Quantex dataset.

The pipeline follows a sequential process: 1. Person and face detection identifies the presence of individuals in the video frames. 2. Gaze classification determines whether detected faces are looking at the child. 3. Voice type classification detects the presence of speech and identifies whether the speaker is the key child, another child, or an adult. 4. Object detection provides insights into the types of objects children interact with, both in social and independent play contexts.

By integrating these modules, our pipeline enables a comprehensive analysis of children's everyday experiences, capturing both social interactions and independent play

In the following sections, we describe each module in detail, including training data, model architecture, and evaluation metrics. A full technical analysis of each algorithm is provided in the Supplementary Material.

Person & Face Detection.

Gaze Classification.

Object Detection.

Voice Detection and Classification.

Results

Presence of Aspects of Social Interaction

Presence of a Person.

Presence of a Face.

Presence of Gaze.

Presence of Language.

Co-occurrence of Aspects of Social Interaction

General Discussion

References

Supplementary Material

YOLO11x: Multi-Class Detection of Persons, Faces, and Objects

In our study, we utilized Ultralytics' YOLO11, the “latest iteration in the Ultralytics YOLO series of real-time object detectors” (Jocher & Qiu, 2024), trained on the COCO dataset. Released in October 2024, YOLO11 introduces architectural improvements such as the C2PSA block (Convolutional Block with Parallel Spatial Attention), which enhances spatial attention within feature maps, allowing the model to focus more precisely on critical areas of an image compared to previous Yolo versions. Additionally, YOLO11 incorporates the C3K2 block, designed to be faster and more efficient, enhancing the overall performance of the feature aggregation process (Khanam & Hussain, 2024). These advancements make the YOLO11 detection model, pretrained on COCO, well-suited for training on our egocentric dataset, which captures dynamic movements from a camera perspective on chest height.

Dataset Preprocessing. Our dataset presents unique challenges due to its egocentric viewpoint, as the body parts of the child wearing the camera frequently appear in the footage. To prevent misclassification, we adopt a dedicated annotation scheme where all individuals in the scene are labeled as “person,” each assigned a unique ID. The key child, who wears the camera, is consistently assigned ID = 1. During preprocessing, we map the key child (ID = 1) into a separate category, “child body parts,” to distinguish their presence from other individuals. Furthermore, we refine the “person” and “face” categories by introducing additional distinctions that standard YOLO models do not inherently make. Specifically, we differentiate between (1) the key child (who wears the camera) and other individuals, (2) adults and children/infants for both full-body detections and faces. These modifications allow for a more precise analysis of social interactions while reducing false positives caused from the key child’s own body while capturing detailed information on the presence and classification of people and objects in the child’s

environment. To address these challenges, our fine-tuned YOLOv1 model is trained to:

- Recognize and differentiate between the key child’s body parts and other individuals.
- Distinguish between adults and children/infants for both full-body detections and faces.
- Identify and classify six specific object categories (toy, book, food, kitchenware, screen, other object) relevant to the child’s interactions.

While our dataset originally included seven object categories, we excluded the “animal” category from training due to its low occurrence (only 19 instances), whereas all other categories had counts exceeding 1,000. As a result, our final model was trained on six object categories.

Dataset Splitting. We started data splitting with a dataset comprising a total of 113799 frames from 80 annotated videos. Prior to splitting this dataset into training, validation, and testing datasets, we analyzed the proportion of frames containing annotated persons, faces and objects versus those without any annotations. Our analysis showed that at least one adult or child (full body or face) was present in 29.75% of the frames and at least one face was visible in 29.75% of the frames.

We applied a stratified split to ensure that each of the training, validation, and testing dataset preserved the original ratio of frames with persons (adult and child/infant), frames with faces (adult and child/infant), frames with objects (6 object classes without animal) to frames without any of the relevant classes. As a result, the final data distribution, displayed in detail in table 1) consisted of 91038 frames in the training dataset, 11379 frames in the validation and 11382 frames in the testing dataset, guaranteeing that the model’s performance evaluation remained accurate to the real-world data distribution.

Training and Convergence. Model training was conducted on a Linux server equipped with an Intel(R) Xeon(R) Silver 4214Y CPU @ 2.20GHz with 48 cores, a Quadro

Table 1

Dataset splits for the YOLO11 person and face detection model trained on the Quantex dataset. The table shows the distribution of annotated persons, faces, and objects in the training, validation, and testing datasets.

Quantex	Percentage	Training	Validation	Testing	Total
Adult Only	6.54	5961	745	746	7452
Child/Infant Only	8.23	7494	936	938	9368
Adult face only	0.15	13	1	3	17
Child/Infant face only	0.12	108	13	14	135
Person & Face	14.84	13506	1688	1689	16883
Book	9.19	8399	1037	1026	10462
Toy	15.23	13879	1721	1737	17337
Kitchenware	2.10	1892	236	266	2394
Screen	3.71	3386	422	410	4218
Food	0.98	900	107	108	1115
Other Object	16.14	14706	1839	1829	18374
Animal	0.02	16	1	2	19
Neither	22.87	20794	2631	2618	26025
Total	100	91037	11377	11385	113799

RTX 8000 GPU and 188 GB of RAM. The model was trained for a total of 86 epochs, taking 200 hours to complete. Training utilized YOLO11's built-in data augmentation, an image size of 640, a batch size of 16, a cosine annealing learning rate scheduler (Loshchilov & Hutter, 2017), and early stopping after 10 epochs without improvement, with a maximum of 200 epochs.

The loss function of the YOLO11 model comprises three main components: Box Loss, Classification Loss, and Distribution Focal Loss (DFL) (Li et al., 2020; Terven, Cordova-Esparza, Ramirez-Pedraza, Chavez-Urbiola, & Romero-Gonzalez, 2024). *Box Loss* quantifies the difference between predicted bounding boxes and ground truth boxes, ensuring precise localization of detected objects by penalizing inaccuracies in position and size. *Classification Loss* evaluates the model's ability to correctly assign detected objects to their respective classes, reducing false positives and false negatives. *Distribution Focal Loss* enhances the model's ability to detect challenging objects, particularly small or partially occluded ones, by refining the localization of bounding box coordinates and emphasizing hard-to-detect instances. Together, these loss components contribute to a more robust and accurate detection model.

During the training process, we observed that all three loss components decreased over time, indicating effective learning and improved performance, as visible in figure 3. A steady decrease in Box Loss indicates that the model is becoming increasingly accurate in localizing persons, faces and objects within frames. Similarly, the steady convergence of the Classification Loss reveals the model's increasing ability to reliably classify the detected object in one of the relevant classes. The decrease in DFL over time indicates that the model is getting better at focusing on and correctly identifying difficult-to-detect persons, faces or objects, which improves its overall detection capabilities. Conclusively, the loss curves show that the model effectively learned to localize and identify the target classes during the training period.

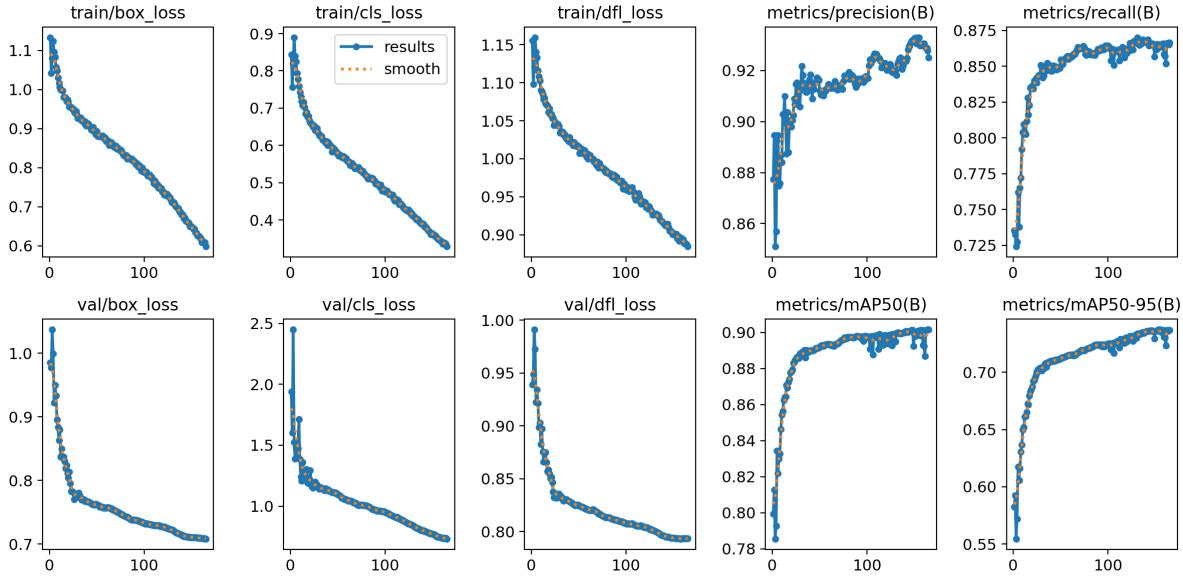


Figure 3. Training and Validation Loss Curves for the YOLO11 person detection model.

Model Evaluation Metrics. The YOLO11 model achieved a precision of 0.92 and a recall of 0.87 on the testing set, resulting in an F1-score of 0.90. These metrics, summarized in table 2 indicate the model’s strong performance in accurately identifying faces while minimizing the number of missed detections and false positives. The precision-recall curve, displayed in figure 4, further illustrates this performance, with the curve remaining close to the top-left corner. This positioning signifies that the model maintains high precision and recall across various thresholds, underscoring its effectiveness in detecting faces with confidence.

Analysis of the confusion matrix reveals that 86% of all faces are correctly identified by the model, corresponding to 1905 true positives, while 234 faces were missed (false negatives). False negatives predominantly occurred in scenarios where faces were in the background, blurred due to motion, or occluded by the child’s body. In such instances, adjacent frames often provided clearer views, aiding in more accurate classification. The model exhibited a false positive rate of approximately 2.72%, with 251 frames incorrectly classified as containing faces when none were present. These false positives were often

Table 2

Evaluation metrics for the YOLO11m detection model trained on the Quantex dataset to detect persons and faces while distinguishing between the key child and other individuals. False Positive Rate and False Negative Rate are given in percentages.

Dataset	Precision	Recall	F1-Score	False Positive Rate	False Negative Rate
Quantex	0	0	0	0	0

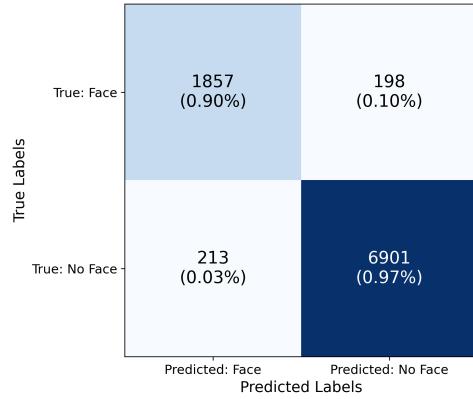
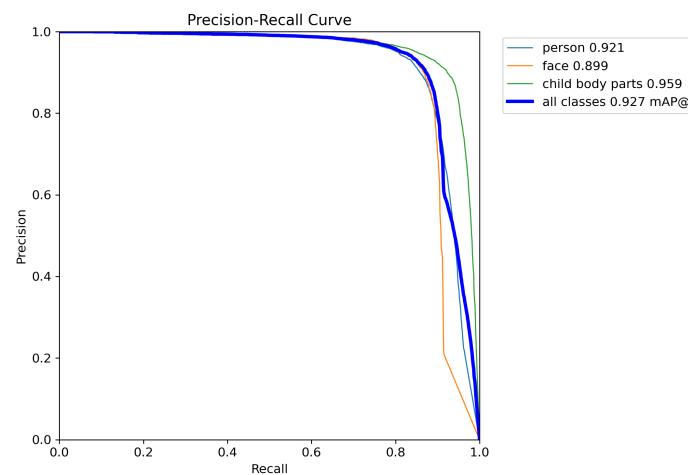
A**B**

Figure 4. **A** - Confusion Matrix for the YOLO11 person and face detection model trained on the Quantex dataset. **B** - Precision-Recall Curve for the YOLO11 person detection model.

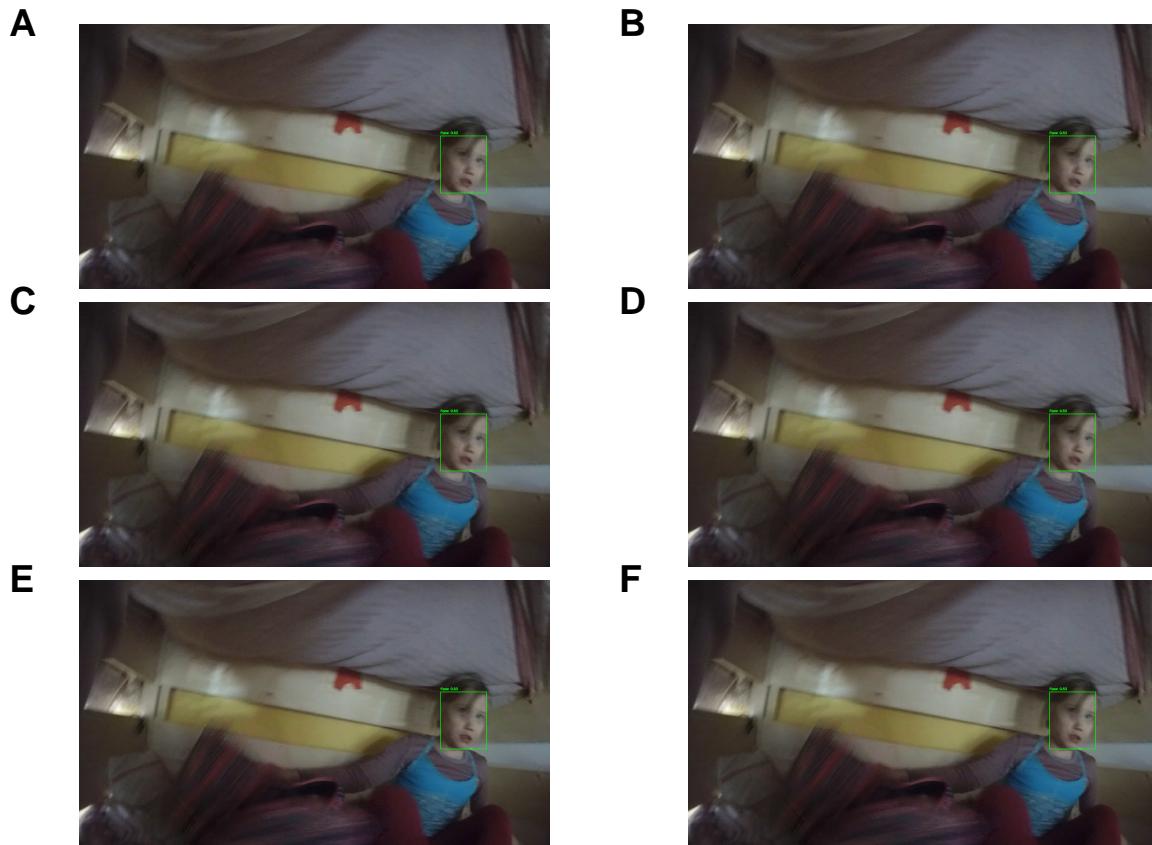


Figure 5. A, B - Examples of True Positives, C, D – Examples of False Negatives, E, F – Examples of False Positives in the YOLO11 person detection model.

attributed to objects or toys resembling facial features. In face detection systems, achieving a balance between false positives and false negatives is crucial. Given this context, a 2.1% false positive rate is generally considered acceptable.

To provide a comprehensive understanding of the model's performance, we have included visual examples of true positives, false positives, and false negatives in figure 5. These images highlight the model's strengths and areas where challenges persist, offering insights into specific scenarios that influence detection accuracy.

Overall, the YOLOv11 model demonstrates robust performance in face detection tasks. However, challenges remain in dynamic scenarios, particularly with partially visible, rotated, or side-view faces. These findings underscore the complexities inherent in

analyzing egocentric video data, where movement and varying perspectives introduce additional challenges.

Yolo11x-cls: Gaze Classification

Selecting an appropriate model for gaze classification in our automated pipeline presented unique challenges due to the egocentric perspective of our dataset. Many gaze estimation methods rely on high-quality eye images, either extracted separately (Zhang, Sugano, Fritz, & Bulling, 2015) or as part of the full face (Zhang, Sugano, Fritz, & Bulling, 2016). However, our dataset often contains blurry or partially occluded faces captured at varying angles, making such approaches not suitable. Additionally, rather than predicting fine-grained gaze direction (e.g., left or right), our focus is on the binary classification of whether a person's gaze is directed toward the child.

(Cheng, Wang, Bao, & Lu, 2021) provides an overview of the challenges and recent advancements in gaze estimation methods, including approaches that incorporate temporal information, such as Gaze360 (Kellnhofer, Recasens, Stent, Matusik, & Torralba, 2019). While these methods improve tracking across frames, they are not the primary focus of our study, as we analyze social interactions on a frame-by-frame basis.

Given these constraints, we opted for a CNN-based approach trained on ground truth annotations. Recent studies (Shah et al., 2022; Zhang et al., 2020; Zhang, Sugano, Fritz, & Bulling, 2019) have explored different CNN based gaze estimation architectures, among which are VGG, ResNet or YOLO architectures. Based on these findings, we implemented a VGG, ResNet, and YOLO-based model for binary classification (gaze directed toward the child or not). In our preliminary tests, Ultralytics' YOLO11 architecture (Jocher & Qiu, 2024) demonstrated the best performance, leading us to select the YOLO11x classification model, pretrained on ImageNet, for our gaze classification task.

Given these constraints, we opted for a CNN-based approach trained on ground truth

annotations. Recent studies (Shah et al., 2022; Zhang et al., 2020, 2019) have explored different CNN-based gaze estimation architectures, including VGG, ResNet, and YOLO. Based on these findings, we implemented and compared three models; VGG, ResNet, and YOLO; for binary classification (gaze directed toward the child or not). In our preliminary tests, Ultralytics' YOLO11 architecture (Jocher & Qiu, 2024) demonstrated the best performance, leading us to select the YOLO11x classification model, pretrained on ImageNet, for our gaze classification task.

The egocentric nature of our video data, recorded from a chest-mounted camera, posed additional challenges such as motion blur and oblique viewing angles, making gaze classification difficult even for human annotators. To address this, we defined gaze as “directed toward the child” when a person’s gaze was oriented toward the child’s face, body, or general direction. This included cases where the person was looking directly at the camera (worn by the child) or slightly upward, estimating the likely position of the child’s head. Each detected face was annotated as either “gaze” (gaze directed at the child) or “no gaze” (gaze directed elsewhere).

YOLO11’s architectural enhancements, such as the C2PSA block, improve the model’s ability to focus on critical regions within an image, making it well-suited for our task. We fine-tuned the pretrained YOLO11x-cls model for gaze classification, enabling it to robustly determine whether individuals in the scene were engaged in visual attention toward the child.

Dataset Splitting. For the gaze classification task, we utilized a subset of the Quantex dataset, focusing specifically on frames containing annotated faces. This subset consisted of cut-out faces extracted from the annotated face bounding boxes, resulting in 20889 frames from 64 annotated videos. An analysis of the gaze labels showed that 21.25% of the faces were annotated as having gaze directed toward the child, resulting in an imbalanced dataset.

Table 3

Dataset splits for the YOLO11 gaze classification model trained on the Quantex dataset. The table shows the total number of frames, as well as the number of frames with gaze and no gaze in the training, validation, and testing datasets after data augmentation of the minority class (Gaze). 'Gaze' indicates frames where the person's gaze is directed towards the child, while 'No Gaze' indicates frames where the person's gaze is not directed towards the child. Ratios are given in percentages.

Quantex	Ratio	Training	Validation	Testing	Total
Gaze	50	13160	1645	1646	16451
No Gaze	50	13160	1645	1646	16451
Total	100.00	26320	3290	3292	32902

Since our dataset was already a subset of the Quantex dataset, downsampling to achieve balance would have led to a loss of valuable data. To address this, we applied data augmentation techniques instead of downsampling, ensuring that the final dataset remained balanced while retaining all available samples. The final dataset comprised 32902 frames, with an equal distribution of 50% gaze-directed and 50% non-gaze frames.

To maintain this balance across all model training stages, we performed stratified dataset splitting which resulted in the data distribution presented in Table 3. Specifically, the training set comprised 26320 frames, the validation set 3290 frames, and the testing set 3292 frames.

Training and Convergence. We trained the gaze classification model on the same linux server that was used for the person and face detection model training. The model

was trained for a total of 165 epochs, taking 283 hours to complete. Similar to the person and face detection model, model training was conducted applying an image size of 640, a batch size of 16, a cosine annealing learning rate scheduler (Loshchilov & Hutter, 2017), and early stopping after 10 epochs without improvement, with a maximum of 200 epochs.

Figure 3 provides an overview of the loss curves for the gaze classification model, showing the Box Loss, Classification Loss, and Distribution Focal Loss (DFL) over the training period. - tbd

Model Evaluation Metrics.

Proximity Heuristic

Proximity between individuals is an important aspect of social interaction. Previous studies have shown that interpersonal distance can provide insights into the nature of relationships and social engagement (Hernández-Heredia, Reyes-Manzano, Flores-Hernández, Ramos-Fernández, & Guzmán-Vargas, 2024; Janssen et al., 2024; Onnela et al., 2014). Since our dataset does not contain explicit proximity labels, we developed a heuristic approach to estimate the distance between the child and other individuals.

Our heuristic relies on the bounding boxes of detected faces, assuming that the size of these bounding boxes provides an implicit cue for proximity. Specifically, we infer that larger bounding boxes correspond to individuals who are closer to the camera, whereas smaller bounding boxes indicate individuals positioned further away. To quantify this relationship, we calculate the area of a detected face bounding box as the product of its width and height. This computed size is then compared to a predefined reference bounding box. For both adults and children/infants, we selected two reference face sizes: one corresponding to a face very close to the camera (within arm's reach) and another representing a face further away (e.g., in the background or outdoors). By normalizing the detected face sizes relative to these reference sizes, we obtain a proximity score ranging

from 0 to 1, where a value of 1 indicates a face extremely close to the camera, and 0 represents a face that is far away.

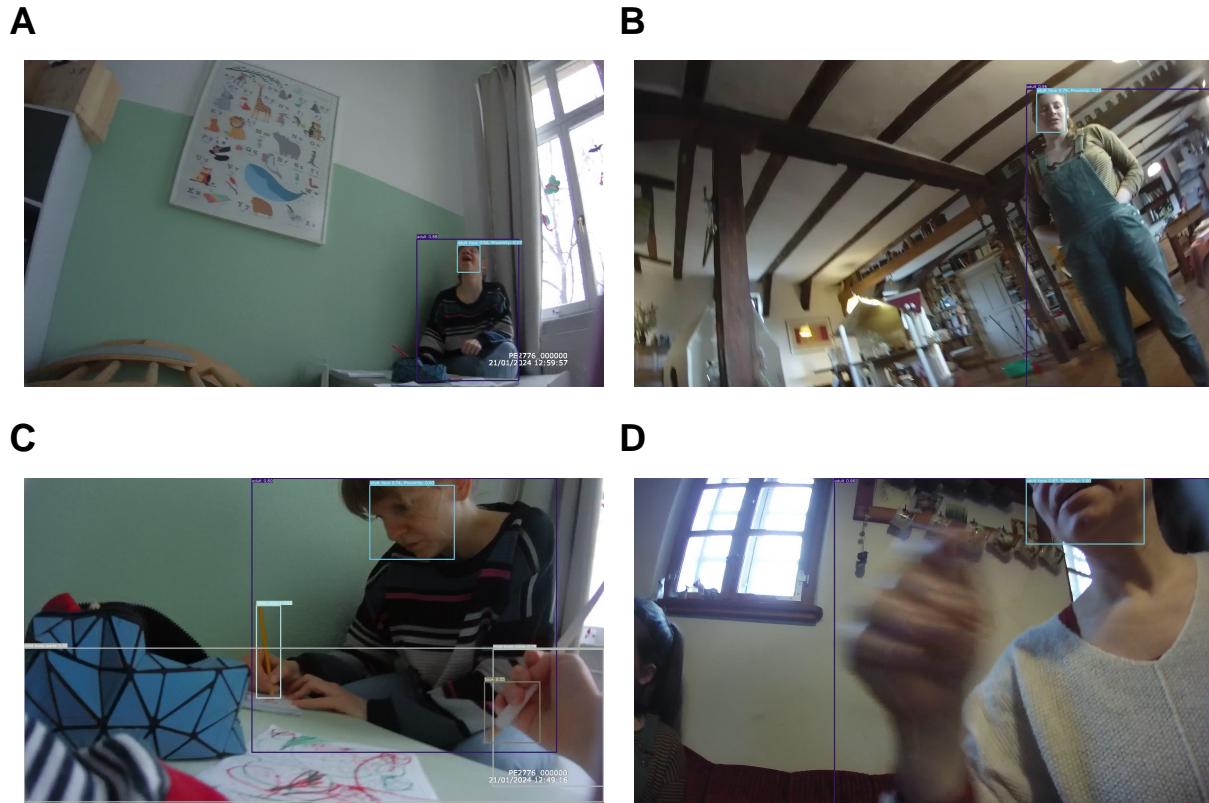


Figure 6. Proximity heuristic examples. Example **A** shows a face far away from the camera (proximity score = 0.07), example **B** depicts a face slightly closer (proximity score = 0.2). Example **C** shows a face quite close to the camera (proximity score = 0.6), and example **D** illustrates a face extremely close to the camera (proximity score = 1).

To better account for the nonlinear relationship between face size and distance, we applied a logarithmic transformation to the bounding box areas. Since face area decreases quadratically as distance increases, a simple linear mapping would make small differences appear too large for close faces and too small for distant ones. The logarithmic transformation helps balance this by compressing large values and stretching smaller ones, creating a more natural and meaningful proximity scale. This approach aligns with how

humans perceive size, as our sensitivity to changes depends on relative differences rather than absolute ones (Stevens & Marks, 2017). By using this scaling, we ensure that proximity estimates remain consistent across different face sizes and distances.

In addition to absolute size, we incorporate a width-to-height ratio as an additional cue for proximity estimation. This ratio was determined separately for adult and child/infant faces based on reference images depicting full, front-facing faces. If a detected face exhibits a significant deviation from the expected aspect ratio, we infer that the face is partially cropped due to extreme proximity to the camera. In such cases, the proximity score is set to 1. An example illustrating this case is shown in Figure ??.

This heuristic provides an interpretable and computationally efficient method for estimating proximity in egocentric recordings, allowing us to examine the spatial dynamics of social interactions without the need for explicit depth information or additional sensors.

Voice Type Classification

Regarding the audio component of our interaction analysis, we aimed to use a model that does not only detect the presence of speech but also distinguishes between the key child wearing the camera and other speakers. This distinction is crucial for understanding the dynamics of the interactions and the role of the key child in the social context. To achieve this, we applied the Voice Type Classifier (Lavechin, Bousbib, Bredin, Dupoux, & Cristia, 2020), an open-source model designed to identify five different voice types: key child, other child, female adult, male adult, and speech in general. The model is based on a convolutional neural network (CNN) architecture and was trained on 260 hours of child-centered recordings across 10 different languages.

Although the Quantex dataset does not include explicit audio labels, we are confident in the model's suitability for our data. Prior testing on a similar labeled dataset which was also collected in our lab, ChildLens, demonstrated that the Voice Type Classifier achieved

an F1 score of 58.1 (reference to ChildLens paper), which is comparable to the F1 score of 57.3 reported on the original training dataset. These results indicate that the model generalizes well to our child-centered recordings, making it a reliable choice for our analysis.

We applied the Voice Type Classifier to the extracted audio data from the Quantex dataset using the code provided by the authors of the voice type classifier (Lavechin, 2020). The model was used to detect the presence of speech and classify the speaker into one of the five predefined categories.

References

- Carpendale, J., & Lewis, C. (2020). *What Makes Us Human: How Minds Develop through Social Interactions* (1st ed.). Routledge. <https://doi.org/10.4324/9781003125105>
- Cheng, Y., Wang, H., Bao, Y., & Lu, F. (2021). Appearance-based Gaze Estimation With Deep Learning: A Review and Benchmark. <https://doi.org/10.48550/ARXIV.2104.12668>
- Dai, S., Bouchet, H., Karsai, M., Chevrot, J.-P., Fleury, E., & Nardy, A. (2022). Longitudinal data collection to follow social network and language development dynamics at preschool. *Scientific Data*, 9(1), 777. <https://doi.org/10.1038/s41597-022-01756-x>
- Hernández-Heredia, T. K., Reyes-Manzano, C. F., Flores-Hernández, D. A., Ramos-Fernández, G., & Guzmán-Vargas, L. (2024). Proximity Sensor for Measuring Social Interaction in a School Environment. *Sensors*, 24(15), 4822. <https://doi.org/10.3390/s24154822>
- Heyes, C. M. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Cambridge (Mass.): Harvard University press.
- Janssen, L. H. C., Verkuil, B., Nedderhoff, A., Van Houtum, L. A. E. M., Wever, M. C. M., & Elzinga, B. M. (2024). Tracking real-time proximity in daily life: A new tool to examine social interactions. *Behavior Research Methods*, 56(7), 7482–7497. <https://doi.org/10.3758/s13428-024-02432-1>
- Jocher, G., & Qiu, J. (2024). *Ultralytics YOLO11*. Retrieved from <https://github.com/ultralytics/ultralytics>
- Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., & Torralba, A. (2019). Gaze360: Physically Unconstrained Gaze Estimation in the Wild. <https://doi.org/10.48550/ARXIV.1910.10088>
- Khanam, R., & Hussain, M. (2024, October 23). YOLOv11: An Overview of the Key Architectural Enhancements. <https://doi.org/10.48550/arXiv.2410.17725>

- Lavechin, M. (2020). *Voice Type Classifier*. Retrieved from
<https://github.com/MarvinLvn/voice-type-classifier>
- Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings.
<https://doi.org/10.48550/ARXIV.2005.12656>
- Lemaignan, S., Edmunds, C. E. R., Senft, E., & Belpaeme, T. (2018). The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PLOS ONE*, 13(10), e0205999.
<https://doi.org/10.1371/journal.pone.0205999>
- Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., ... Yang, J. (2020, June 8). Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. <https://doi.org/10.48550/arXiv.2006.04388>
- Loshchilov, I., & Hutter, F. (2017, May 3). SGDR: Stochastic Gradient Descent with Warm Restarts. <https://doi.org/10.48550/arXiv.1608.03983>
- Onnela, J.-P., Waber, B. N., Pentland, A., Schnorf, S., & Lazer, D. (2014). Using sociometers to quantify social interaction patterns. *Scientific Reports*, 4(1), 5604.
<https://doi.org/10.1038/srep05604>
- Piaget, J. (1964). Part I: Cognitive development in children: Piaget development and learning. *Journal of Research in Science Teaching*, 2(3), 176–186.
<https://doi.org/10.1002/tea.3660020306>
- Rogoff, B., Dahl, A., & Callanan, M. (2018). The importance of understanding children's lived experience. *Developmental Review*, 50, 5–15.
<https://doi.org/10.1016/j.dr.2018.05.006>
- Rossano, F., Terwilliger, J., Bangerter, A., Genty, E., Heesen, R., & Zuberbühler, K. (2022). How 2- and 4-year-old children coordinate social interactions with peers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1859), 20210100. <https://doi.org/10.1098/rstb.2021.0100>

- Shah, S. M., Sun, Z., Zaman, K., Hussain, A., Shoaib, M., & Pei, L. (2022). A Driver Gaze Estimation Method Based on Deep Learning. *Sensors*, 22(10), 3959.
<https://doi.org/10.3390/s22103959>
- Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The Developing Infant Creates a Curriculum for Statistical Learning. *Trends in Cognitive Sciences*, 22(4), 325–336.
<https://doi.org/10.1016/j.tics.2018.02.004>
- Stevens, S. S., & Marks, L. E. (2017). *Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects* (1st ed.). Routledge.
<https://doi.org/10.4324/9781315127675>
- Terven, J., Cordova-Esparza, D. M., Ramirez-Pedraza, A., Chavez-Urbiola, E. A., & Romero-Gonzalez, J. A. (2024, October 12). Loss Functions and Metrics in Deep Learning. <https://doi.org/10.48550/arXiv.2307.02694>
- Tomasello, M. (2009). *Cultural Origins of Human Cognition*. Cambridge: Harvard University Press.
- Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.
- Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., & Hilliges, O. (2020). ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation. <https://doi.org/10.48550/ARXIV.2007.15837>
- Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2015). Appearance-based gaze estimation in the wild. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4511–4520. Boston, MA, USA: IEEE.
<https://doi.org/10.1109/CVPR.2015.7299081>
- Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2016). *It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation*.
<https://doi.org/10.48550/ARXIV.1611.08860>
- Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2019). MPIIGaze: Real-World Dataset

and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1), 162–175.

<https://doi.org/10.1109/TPAMI.2017.2778103>

Appendix