Automatic Detection of the Visual Gaze Components of Joint Attention in Naturalistic, Observational Data

Miranda Dickerman, Anshul Gupta, Samy Tafasca, Xiaocheng Zhang, Jean-Marc Odobez, and Sabine Stoll

Early language acquisition depends on interaction with others, with joint attention playing a key role. Joint attention is an interactional framework characterized by shared focus on and interaction about some third object or activity of interest (Bakeman & Adamson 1984). Children become capable of jointly attending around nine months of age (Trevarthen & Hubley 1978; Callaghan et al. 2011). Joint attention has been consistently linked to vocabulary development (e.g., Tomasello & Farrar 1986; Akhtar et al. 1991) and broader language related skills (e.g., Carpenter et al. 1998; Adamson et al. 2019). Joint attention itself varies in form in several ways: it is comprised of different social components, such as the individuals involved (e.g., mother, father, or peers), and different interactional components, such as speech, gaze, and gesture. These components in turn impact the ability of young children to learn from joint attention (Suarez-Rivera et al. 2019; Abney et al. 2020). Gaze behavior is a particularly key aspect of joint attention; variance in gaze behavior has been linked to different learning outcomes with the use of eye-tracking devices (Abney et al. 2020). Here we take a first step into tailoring automatic gaze estimation tools to naturalistic child language acquisition videos in the absence of specialized eye-tracking equipment. We apply computational gaze estimation models to a naturalistic, observational language acquisition corpus, featuring children between the ages of 2;0 and 4;0 learning Tuatschin-Romansh. Our objectives are (1) to examine fine-grained gaze behavior during naturalistic joint attention in the absence of eye-tracking equipment, (2) to assess the feasibility of automatically estimating these components using machine learning architectures, and (3) to examine if gaze-point estimates are informative about joint attention in our dataset.

---

*M. Dickerman, Institute for the Interdisciplinary Study of Language Evolution (ISLE Institute), University of Zürich, miranda.dickerman@uzh.ch; A. Gupta, Idiap Research Institute (Idiap), École Polytechnique Fédérale de Lausanne (EPFL); S. Tafasca, Idiap, EPFL; X. Zhang, EPFL; J.M. Odobez, Idiap, EPFL; S. Stoll, ISLE Institute, University of Zürich.
The authors warmly thank their research assistants from the University of Languages and International Studies, Vietnam National University. This work was supported by the NCCR Evolving Language, Swiss National Science Foundation (no. 51NF40_180888), and the AI4Autism project (Digital Phenotyping of Autism Spectrum Disorders in Children, CRSII5 202235/1) of the SNSF-Sinergia programme.

 Proceedings of the 49th annual Boston University Conference on Language Development, ed. Aditya Yedetore et al., 199-212. Somerville, MA: Cascadilla Press.

Joint attention is considered foundational for language learning because it acts as a connective mechanism between language and real-world referents; the language produced during interaction about some focal referent can be clearly connected to that referent (Baldwin 1995). Here referred to as joint attention, various terms, including coordinated joint attention (Bakeman & Adamson 1984; Carpenter et al. 1998; Mastin & Vogt 2016), triadic interaction (Little et al. 2016), and joint engagement (Bard et al. 2021; Schatz et al. 2022), have been used to refer to essentially the same framework of two – in some definitions, two or more – individuals interacting about some third focal object of attention. In a Western context, a prototypical joint attentional frame involves a child, their parent, and some toy (Bard et al. 2021). There is some evidence that joint attention may be less frequent in some cultural contexts (Brown 2011; Taverna et al. 2024) and that this reduced frequency affects the relevance of joint attention for word learning (Mastin & Vogt 2016). This remains a point of contention, as other work has found that the amount of joint attention does not differ across diverse cultural contexts (Bard et al. 2021). Research on autism also provides evidence for the relevance of joint attention to language learning: children with autism are consistently behind their typically developing peers in their use of joint attention (Adamson et al. 2009), which ultimately affects their vocabulary skills (Adamson et al. 2019).

The composition of joint attention may also support or hinder its role as a road-map to word meaning and syntactic development. Studies using wearable eye-tracking devices have provided fine-grained insight into gaze behavior during joint attention. For example, Yu et al. (2019) found that infants' sustained attention during joint attention strongly predicted vocabulary size three and six months after the recorded play session. Abney et al. (2020) found that a combination of parent triadic gaze (alternating gaze between their child's face and the focal toy), in conjunction with infant focus on the toy, correlated with future vocabulary size, but other gaze components measured within the study (such as simultaneous shared attention on the object) did not. Deconstructing joint attention into fine-grained gaze components has only been possible thus far with the use of wearable eye-tracking devices. Wearing bulky equipment inherently alters participants' experiences, which may have an effect on the eventual research outcome: in other words, it is unknown if the correlations observed in these studies contribute to learning processes outside of the lab.

Most studies examining joint attention take place in environments which are controlled in some way, either through the identities of interactants, the task given to participants, or the physical setting of the study. For example, Abney et al. (2020) placed mother-infant dyads across a table from each other in a laboratory, with three toys in front of them, and instructed them to play together; this is referred to as "naturalistic free-flowing interaction" (Abney et al. 2020). Mother-infant dyads are commonly studied, both in early work (e.g. Bakeman & Adamson 1984; Tomasello & Farrar 1986) and in contemporary approaches (e.g., Suarez-Rivera et al. 2022) despite a recent finding that joint attention with fathers may be a stronger contributor to vocabulary size than joint attention with mothers (Atamán-

Devrim et al. 2023). Setting is less often controlled; although many influential studies on joint attention take place in the laboratory (e.g., Akhtar et al. 1991; Carpenter et al. 1998), home environments are also commonly studied (Tomasello & Todd 1983; Bakeman & Adamson 1984; Suarez-Rivera et al. 2022). Joint attention has also been studied in a relatively wide range of cultural and socio-economic contexts (Callaghan et al. 2011; Mastin & Vogt 2016; Abels 2020; Bard et al. 2021). Despite this, prototypical joint attention research focuses on dyadic interaction with mothers, in a semi-prompted setting, often one of free play. Different contexts are known to lead to different research outcomes (Tamis-LeMonda et al. 2017) and as such, the strong focus on a particular type of joint attention limits the generalization of research findings. Not much is known about joint attention in unprompted settings, without controlling for the identities of existing interactants; furthermore, nothing is known about fine-grained, second-by-second gaze behavior during joint attention in the absence of specialized eye-tracking equipment.

Research on behavior, including joint attention, is hindered by the need for manual coding. As a result, studies usually analyze small samples and generalize their findings to the larger population. This practice has contributed to the replication crisis in the psychological sciences (Shrout & Rodgers 2018). In order to test if statistically significant findings are generalizable to wider populations, 'big data' approaches are recommended. This involves using all available data, which substantially increases the amount of hand-annotations required. Advancements in artificial intelligence offer promising solutions to alleviate this workload. Recent advances in computer vision have rapidly improved computational image processing and gaze-target detection. Gaze-target detection – estimating where a person in a static image is likely to be looking – is a topic of broad interest. In recent years, several architectures have been developed for gaze-target estimation (e.g., Chong et al. 2020; Tafasca et al. 2024), and large datasets have been created to train neural networks for this task (e.g., Tafasca et al. 2023; Chong et al. 2020). More recently, the gaze-target task has been expanded to socially relevant applications, such as estimating simultaneous shared attention on some object (e.g., Sumer et al. 2020; Li et al. 2023; Nakatani et al. 2023) or looks to others' faces (Gupta et al. 2024). However, gaze-target estimation has not yet been applied to naturalistic data nor to research in child language acquisition. This is due to two primary reasons. First, implementing gaze detection architectures is a highly technical undertaking. While Erel et al. (2023) is one recent user-friendly adoption of advances in gaze estimation technology, it is limited to direction estimation; gaze-target estimation algorithms have no analogous tool. Second, there is a lack of appropriate datasets. As Tafasca et al. (2023) points out, most gaze estimation datasets feature adults: models trained on primarily adult data are known to perform significantly worse on data featuring children (Sciortino et al. 2017).

In this study, we present a first attempt to adapt gaze-target estimation models to naturalistic child language data. We extract and micro-code samples from a naturalistic corpus of first language acquisition (Tuatschin-Romansh) in order to examine the feasibility of applying automated annotation procedures

to the dataset. We apply publicly trained gaze-target estimation architectures to our naturalistic dataset and compare manual coding work to automated outputs. Finally, we conduct a frame-wise investigation into whether gaze-point estimates can be informative about joint attention in our newly constructed dataset.

1. Method

1.1. Data

Data were extracted from an unpublished corpus of Romansh-Tuatschin (Mazara et al. Unpublished), a dialect of the Sursilvan variety of Romansh spoken in the Grisons region in Switzerland (Maurer-Cecchini 2021). The corpus consists of longitudinal, naturalistic audiovisual recordings of six children between the ages of 2;0 and 4;3. Each child was recorded for approximately 4.5 hours in monthly intervals for either one or two years. Recordings were carried out by the children's parents. The recordings are unprompted and feature the target children doing typical activities at home, such as playing alone, with siblings, with their parents, or with grandparents. The corpus is extremely visually diverse, with the camera placed in different locations around the children's homes, capturing interactants near and far from the camera, sometimes only partially visible.

Two datasets were constructed from samples randomly drawn from this corpus – a testing dataset and a training dataset. This split allowed us to first fine-tune a gaze-target estimation model on the training dataset, and then evaluate its performance on the testing dataset. Fine-tuning is particularly important given the domain gap between our dataset and publicly available datasets: while public datasets tend to focus on adults, our dataset is captured with a larger camera field-of-view, and is more centered on household settings, compared to the more diverse contexts in public datasets. Clips were sampled without replacement and consisted of 10-second-long video clips, sampled from the dataset at a frame rate of 10 frames-per-second, for a total of 100 frames per video for annotation.

1.1.1. Testing dataset

The testing dataset comprised 20 video-clips, or 2,000 frames, and was fully annotated four times by four independent coders, with Coder A treated as ground truth for model evaluations. The data for this dataset was hand-picked from the Tuatschin-Romansch corpus in the following way: clips were randomly sampled from the corpus, and then reviewed for high visual quality. Clips that poorly showed study participants, had camera motion in them, or were otherwise judged to be visually poor were excluded from the testing dataset. The testing dataset was not chosen for ecological validity but rather to examine model estimates given 'good' use-cases from the naturalistic data.

1.1.2. Training dataset

The training dataset comprised 519 video-clips, or 51,900 frames. Clips were sampled randomly from the corpus. Since some randomly sampled clips lacked participants, additional clips were later added to the dataset to compensate. There are a total of 19 clips in which no person is visible at any time, and 500 clips in which at least one person appears at any given time. This dataset was annotated by a team of 14 independent coders.

1.2. Annotation

1.2.1. Manual annotation

Annotation was done using the tool Labelbox (Labelbox 2024). The coding procedure consisted of two 'passes' through each video. The first pass involved coding frame-wise elements, with coders working on one participant at a time. They marked the location of participants' heads in the video-clip by drawing a *head crop* around their head, and marked gaze-point by estimating where in the video participants were looking (see Figure 1). Looking off-screen was coded as *out-of-frame*; participants with heads occluded from the camera's view were coded as *occluded head-crop*; similarly, if gaze-targets were within the camera's view but occluded from sight, this was *occluded gaze-point*. *Eyes closed* was coded if no gaze-target could be selected due to closed eyes. *Looking-at-each-other*, when two participants looked at each other simultaneously, and *shared attention*, when two or more participants looked at the same object or activity simultaneously, were coded frame-wise. Joint attention was coded in the second pass. Joint attention was defined to have a minimum duration of 20 frames (2 seconds) following Gabouer & Bortfeld (2021), and was coded when two or more people were interacting about some shared focal object, activity, or, in some cases, person. All participants in all videos were annotated. Coders underwent an in-person training session on the annotation protocol and completed practice annotations. They received feedback on the practice annotations, and corrected them as instructed before continuing to the annotation task. All videos in the training dataset were coded once. 45% of videos in the training dataset were reviewed and corrected by a secondary experienced coder.

Figure 1: Coding task and model outputs. Images are from the ChildPlay dataset (Tafasca et al. 2023). (A) demonstrates the coding task of drawing a head crop and estimating gaze-point. (B) shows Sharingan model outputs (Tafasca et al. 2024) which estimate gaze-points. (C) shows MTGS model outputs (Gupta et al. 2024) which estimates gaze-points and social gaze components, Looking at Head (LAH) and Looking at Each Other (LAEO).

1.2.2. Automated annotation

We applied two publicly trained gaze-target estimation models to the testing dataset described above. Visualizations of the applied models and their respective outputs are available in Figure 1 (B) and (C). Sharingan is a transformer-based architecture for gaze following that predicts the gaze target for multiple people in the scene (Tafasca et al. 2024). MTGS estimates gaze-targets and social gaze components within a single architecture (Gupta et al. 2024) while being multi-person and temporal. Both models output per-person gaze heat-maps, and the point of maximum intensity is selected as the gaze point.

Estimates from the Sharingan architecture (Tafasca et al. 2024) were generated using a model pre-trained on GazeFollow (Recasens et al. 2017) and fine-tuned on VideoAttentionTarget (Chong et al. 2020). The publicly-trained MTGS model was also pre-trained on GazeFollow, but was fine-tuned on VSGaze, which, unlike VideoAttentionTarget, contains data of children interacting (Gupta et al. 2024). Our fine-tuned MTGS model built upon the publicly-trained version by undergoing additional training on our training dataset. It was initialized with pre-trained weights and iteratively updated using 450 clips from our training dataset; the remaining 50 clips were reserved for validating and selecting the best model. The metric used for model selection was distance, which is the average Euclidean distance between ground truth and predicted gaze points with x- and y- axes scaled to [0, 1]. The model was fine-tuned for 20 epochs. We used an AdamW optimizer, a learning rate of 3×10^{-5}, and a cosine annealing schedule with warm restarts. Stochastic weight averaging was applied starting from epoch 12 with a base learning rate of 1×10^{-5}. The model was fine-tuned without freezing any modules to maximize adaptation to our dataset. The resulting fine-tuned model was applied to our testing dataset, enabling us to compare outputs of all three models.

1.2.3. Evaluation metrics

Inter-coder reliability is essential not only to validate reliability in hand-coded data but also to establish reasonable expectations for model estimates. We calculated the intraclass correlation coefficient (ICC) following Koo & Li (2016), separating x- and y- coordinates due to differences in their possible range. Four independent coders annotated all 2,000 frames in the testing dataset. Coder A was treated as the ground-truth. ICC assumes data-point independence; in our datasets, temporally adjacent data-points are more similar than non-adjacent data-points. To address interdependency between data-points, one set of gaze-point coordinates was sampled per individual in each video clip. This resulted in 48 data-points, above the recommended threshold of 30 (Koo & Li 2016).

The distance metric, a measure typically used to assess automatic gaze-estimation accuracy, was also calculated. The distance metric is the mean of all distances between gaze-point estimates relative to the ground truth, with x and y axes scaled to [0, 1]. Coder A was again treated as the ground truth. This metric also faces the issue of inter-dependent data-points, but in order to follow the standard, distance was calculated by aggregating all data-points. We also calculated a separate distance metric using the random sample extracted for ICC calculation, to allow for comparison between independent data-points.

Both metrics above were calculated to assess coding reliability between human coders. We also employed them to assess the performance of the three computational models applied to this dataset: publicly trained models Sharingan and MTGS, and the fine-tuned MTGS model. For both metrics, only data-points coded by all four coders were included in the calculation, as is standard for the ICC metric. Some gaze-points were flagged with gaze-point exceptions – for example, *eyes-closed*, *obscured gaze-point* and so on. Gaze-points coded with the exceptions *eyes closed, obscured head crop, obscured gaze-point,* and *out-of-frame* were removed from the pool prior to random sampling.

To examine the distribution of social gaze cues relative to joint attention, we conducted a descriptive, frame-wise analysis. Metrics included *LAH*, Looking at Heads, *LAEO*, Looking at Each Other, and *SA*, Shared Attention (simultaneous shared attention to the same object). Shared attention was coded by annotators; the other metrics were calculated by post-processing gaze-point estimates relative to coded head crops, allowing for comparison of model estimates to ground-truth. Only the MTGS model directly predicted these social gaze components; as such, for comparison across all models, we additionally estimated shared attention by calculating the pairwise distance between gaze-points in scenes with only two participants. We compared this metric of gaze-point distance to scenes with and without joint attention. This was done to examine if gaze information could be informative about joint attention in the dataset.

2. Results

2.1. Descriptive statistics

Of the 51,900 frames in the training dataset, 49,355 contained at least one person. 113,776 head-crops were coded, of which 88,855 had a gaze-point estimate. Missing gaze-point estimates were largely coded by one or more gaze-exceptions: eyes closed (2,310), out-of-frame (12,308), obscured head-crop (6,605), and obscured gaze-point (17,471). Shared attention was coded in 13,378 participants, eye contact in 3,460 participants, and joint attention in 22,487 participants.

In the testing dataset, all 2,000 frames contained at least one person. A total of 4,931 head-crops were coded, of which 4,014 had a gaze-point estimate. Gaze exceptions coded were eyes closed (109), out-of-frame (318), obscured head-crop (346), and obscured gaze-point (480). Shared attention was coded in 1,418 participants, eye contact in 323 participants, and joint attention in 2,346 participants.

2.2. Inter-coder reliability metrics

Manual coder reliability was assessed with a two-way random effects model for absolute agreement, yielding excellent agreement for both x- (0.91, 95% CI [0.88, 0.95]) and y-coordinates (0.96, 95% CI [0.94, 0.98]). Automated gaze-target estimation reliability was assessed for each model against ground truth (Coder A) using a two-way mixed effects model for absolute agreement. Manual coders were also individually assessed against ground truth with two-way random effects models for absolute agreement. ICC reliability may be interpreted as: $<$ 0.5: *poor*, 0.5 - 0.75: *moderate*, 0.75 - 0.9: *good*, and $>$ 0.9: *excellent*. Manually coded data generally shows excellent reliability, and model-coded data shows moderate to good reliability. Both ICC and the distance metric are available in Table 1. In both metrics, the model which performs best is the fine-tuned MTGS model, and human coders outperform all automatic coding methods. Comparisons are in Table 1.

Table 1. Metrics comparing manual and automated gaze-point estimates.

	Distance	Sampled Distance	ICC_x	ICC_y
	μ(σ) (n=3,116)	*μ(σ) (n=48)*	ICC_x [95% CI] (*n=48*)	ICC_y [95% CI] *(n=48)*
Coder B	0.047 (0.074)	0.031 (0.031)	0.917 [0.857, 0.952]	0.962 [0.919, 0.981]
Coder C	0.048 (0.084)	0.050 (0.088)	0.884 [0.802, 0.933]	0.964 [0.936, 0.980]
Coder D	0.049 (0.081)	0.057 (0.107)	0.940 [0.896, 0.966]	0.955 [0.922, 0.975]
Sharingan	0.124 (0.135)	0.134 (0.170)	0.645 [0.444, 0.784]	0.684 [0.498, 0.809]
MTGS	0.112 (0.112)	0.124 (0.131)	0.832 [0.719, 0.902]	0.677 [0.489, 0.805]
Fine-tuned MTGS	0.098 (0.111)	0.090 (0.086)	0.876 [0.790, 0.929]	0.876 [0.789, 0.929]

To compare the fine-tuned MTGS model to the publicly trained version, we assessed its direct prediction of categorical social gaze behaviors. Precision, recall, and F1 metrics for both models are available in Table 2.

Table 2. Metrics comparing publicly-trained and fine-tuned MTGS.

	Shared Attention		Looking at Heads		Looking at Each Other	
	Publicly trained	Fine-tuned	Publicly trained	Fine-tuned	Publicly trained	Fine-tuned
Precision	0.329	**0.407**	0.608	**0.699**	**0.902**	0.852
Recall	**0.793**	0.711	0.432	**0.451**	**0.372**	0.351
F1	0.465	**0.518**	0.505	**0.548**	**0.527**	0.497

2.3. Distribution of social gaze components relative to joint attention

In the training dataset, we assessed the frame-wise frequency of each social gaze component in participants coded with and without joint attention. Joint attention occurs in 22,487 participants in the dataset. Of these, 2,990 participants (13.3%) were Looking at Heads (LAH), 824 (3.7%) were Looking at Each Other (LAEO), and 4,158 (18.5%) were Sharing Attention (SA). Joint attention does not occur in 91,296 participants in the dataset. Of these, 10,231 participants (11.2%) were LAH, 2,636 (2.9%) were LAEO, and 9,220 (10.1%) were SA. Significance testing was not carried out due to interrelated data-points.

In the testing dataset, joint attention occurs in 2,346 participants. Of these, 466 (19.9%) were LAH, 219 (9.3%) were LAEO, and 965 (41.1%) were SA. Joint

attention does not occur in 2,588 participants. Of these, 368 (14.2%) were LAH, 104 (4.0%) were LAEO, and 453 were (17.5%) SA. To compare all models, we also calculated the distance between participants' gaze-points during and outside of joint attention, as a proxy to shared attention. We restricted the sample to only include frames with two people coded. Distributions are shown in Figure 2.

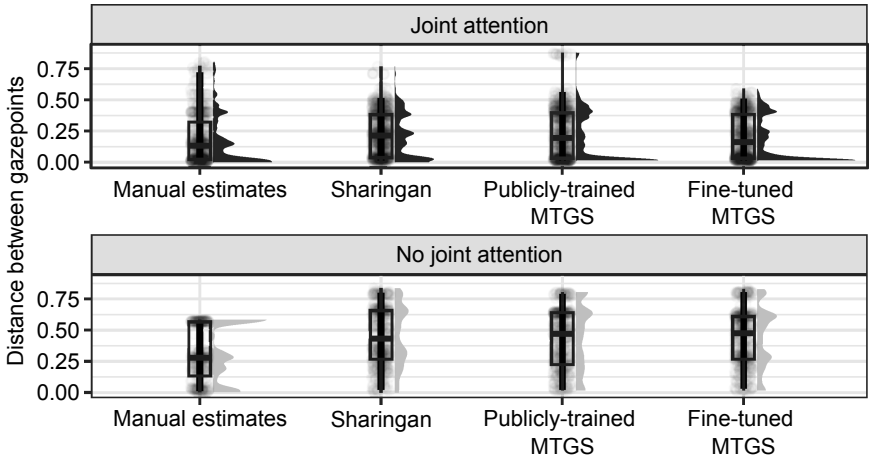

Figure 2. Distance between gaze-points in frames with only two people in the testing dataset. We compare four coding methods; manual estimates, Sharingan estimates, and both publicly-trained and fine-tuned MTGS estimates. We observe similar distributions across all coding methods during and outside of joint attentional frames.

3. Discussion

In this paper we applied automatic gaze-target estimation architectures to naturalistic data of language acquisition in order to evaluate their performance. Inter-coder reliability metrics demonstrate differences in performance between models, with the fine-tuning model generally performing the most like human coders in gaze-target estimation. Differences between Sharingan and MTGS performance may be due to their training datasets; MTGS was trained on a dataset containing videos of adults and children, while Sharingan was only trained on adult data. We also conclude that, while manual coding outperforms all models evaluated on our testing set, automatic estimates still perform within reliability standards deemed acceptable by the scientific community. We also observe that fine-tuning on like data generally improves model performance on our testing dataset, with the exception of some social gaze metrics. Despite these promising findings, it is unknown whether model estimates may be biased in some way not detectable by this analysis. For example, gaze-point estimates produced by models may be more likely to fall nearby hands handling objects than the manual gaze-point estimates, or may

reflect other biases present in the training data. We recommend caution: before drawing conclusions about gaze behavior from automatic estimates, it is important to establish in what context biased estimates may occur.

We show that calculating pairwise distance from automatically extracted gaze-points can be probabilistically informative about joint attention (see Figure 2). Even though model outputs do not precisely reflect manual estimations, similar gaze-point distance distributions relative to joint attention are evident across coding methodologies. This suggests that, even though precision and accuracy of our computational methods still warrant improvement on a frame-wise level, application of gaze-estimation models may already be informative for temporally broader questions, such as if joint attention is likely to exist within a video. Figure 2 also highlights the improvement of the fine-tuned MTGS model; after fine-tuning, outliers disappear from the distribution, and the clustering differences between joint attention conditions are clearer. This demonstrates that building fine-tuning datasets – a smaller and less daunting task than constructing full-sized training datasets– is a practical way to improve the annotation reliability of gaze-target estimation.

As this was a first attempt to automatically estimate gaze-targets in naturalistic data, we curated our testing dataset by choosing visually optimal videos. This was done so as to maximize model performance given high quality video data, which is not consistently possible to collect in naturalistic settings but may be frequent in other settings, such as laboratory studies. Our results therefore are not generalizable to the dataset as a whole, but may be indicative of model performance in similarly controlled settings. Naturalistic data presents inherent challenges due to its complexity and lack of structure. Even when manually coding the videos selected for high visual quality, some images were difficult to code; this includes instances when children were in motion, partially obscured by something in the camera's field, or facing away from the camera. Despite these considerations, it is promising that our results demonstrate good performance on hand-selected clips, and that these results improve after fine-tuning on the naturalistic training dataset. More work is necessary to determine the limitations of automatic gaze-point estimation relative to video quality.

4. Conclusion

This study was the first to apply automatic gaze-target estimation, based on third-party-view video, to a language acquisition corpus. We find that gaze-point estimates provide probabilistic insights into joint attention in our dataset, with publicly-trained gaze estimation architectures reflecting this relationship. While our results are not indicative of precise automatic gaze-target estimation, they do demonstrate that automatic gaze-target estimates meet reliability criteria and are informative about joint attention in our dataset. Ultimately, we suggest that automated gaze-target estimation is a promising new tool for accessing previously untapped data-types and environments in psychological research.

References

Abels, Monika. 2020. Triadic interaction and gestural communication: Hierarchical and child-centered interactions of rural and urban Gujarati (Indian) caregivers and 9-month-old infants. *Developmental Psychology* 56(10). 1817.

Abney, Drew H, Sumarga H Suanda, Linda B Smith & Chen Yu. 2020. What are the building blocks of parent–infant coordinated attention in free-flowing interaction? *Infancy* 25(6). 871–887.

Adamson, Lauren B, Roger Bakeman, Deborah F Deckner & MaryAnn Romski. 2009. Joint engagement and the emergence of language in children with autism and Down syndrome. *Journal of Autism and Developmental Disorders* 39. 84–96.

Adamson, Lauren B, Roger Bakeman, Katharine Suma & Diana L Robins. 2019. An expanded view of joint attention: Skill, engagement, and language in typical development and autism. *Child Development* 90(1). e1–e18.

Akhtar, Nameera, Frances Dunham & Philip J Dunham. 1991. Directive interactions and early vocabulary development: The role of joint attentional focus. *Journal of Child Language* 18(1). 41–49.

Ataman-Devrim, Merve, Elizabeth Nixon & Jean Quigley. 2023. Joint attention episodes during interactions with fathers but not mothers at age 2 years is associated with expressive language at 3 years. *Journal of Experimental Child Psychology* 226. 105569.

Bakeman, Roger & Lauren B Adamson. 1984. Coordinating attention to people and objects in mother-infant and peer-infant interaction. *Child Development* 1278–1289.

Baldwin, Dare A. 1995. Understanding the link between joint attention and language. In *Joint Attention*, 131–158. Psychology Press.

Bard, Kim A, Heidi Keller, Kirsty M Ross, Barry Hewlett, Lauren Butler, Sarah T Boysen & Tetsuro Matsuzawa. 2021. Joint attention in human and chimpanzee infants in varied socio-ecological contexts. *Monographs of the Society for Research in Child Development* 86(4). 7–217.

Brown, Penelope. 2011. The cultural organization of attention. In *The Handbook of Language Socialization*, 29–55. Wiley Online Library.

Callaghan, Tara, Henrike Moll, Hannes Rakoczy, Felix Warneken, Ulf Liszkowski, Tanya Behne, Michael Tomasello & W Andrew Collins. 2011. Early social cognition in three cultural contexts. *Monographs of the Society for Research in Child Development* i–142.

Carpenter, Malinda, Katherine Nagell, Michael Tomasello, George Butterworth & Chris Moore. 1998. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development* i–174.

Chong, Eunji, Yongxin Wang, Nataniel Ruiz & James M Rehg. 2020. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5396–5406.

Erel, Yotam, Katherine Adams Shannon, Junyi Chu, Kim Scott, Melissa Kline Struhl, Peng Cao, Xincheng Tan, Peter Hart, Gal Raz, Sabrina Piccolo et al. 2023. iCatcher+: Robust and automated annotation of infants' and young children's gaze behavior from videos collected in laboratory, field, and online studies. *Advances in Methods and Practices in Psychological Science* 6(2). 25152459221147250.

Gabouer, Allison & Heather Bortfeld. 2021. Revisiting how we operationalize joint attention. *Infant Behavior and Development* 63. 101566.

Gupta, Anshul, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard & Jean-Marc Odobez. 2024. MTGS: A Novel Framework for Multi-Person Temporal Gaze Following and Social Gaze Prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, (NeurIPS).

Koo, Terry K & Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15(2). 155–163.

Labelbox. 2024. Labelbox. Available: Online. https://labelbox.com/.

Li, Peitong, Hui Lu, Ronald W Poppe & Albert Ali Salah. 2023. Automated detection of joint attention and mutual gaze in free play parent-child interactions. In *Companion Publication of the 25th International Conference on Multimodal Interaction*, 374–382.

Little, Emily E, Leslie J Carver & Cristine H Legare. 2016. Cultural variation in triadic infant–caregiver object exploration. *Child Development* 87(4). 1130–1145.

Mastin, J Douglas & Paul Vogt. 2016. Infant engagement and early vocabulary development: a naturalistic observation study of Mozambican infants from 1; 1 to 2; 1. *Journal of Child Language* 43(2). 235–264.

Maurer-Cecchini, Philippe. 2021. *A grammar of Tuatschin: A Sursilvan Romansh dialect (Volume 3)*. Language Science Press.

Mazara, Jekaterina, Géraldine Walther, Benoit Sagot, Claudia Cathomas, Michele Loporcaro & Sabine Stoll. Unpublished. Audiovisual Longitudinal Corpus of 6 Children Learning Romansh Tuatschin.

Nakatani, Chihiro, Hiroaki Kawashima & Norimichi Ukita. 2023. Interaction-aware joint attention estimation using people attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10224–10233.

Recasens, Adria, Carl Vondrick, Aditya Khosla & Antonio Torralba. 2017. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, 1435–1443.

Schatz, Jacob L, Catalina Suarez-Rivera, Brianna E Kaplan & Catherine S Tamis-LeMonda. 2022. Infants' object interactions are long and complex during everyday joint engagement. *Developmental Science* 25(4). e13239.

Sciortino, Giuseppa, Giovanni Maria Farinella, Sebastiano Battiato, Marco Leo & Cosimo Distante. 2017. On the estimation of children's poses. In *Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part II 19*, 410–421. Springer.

Shrout, Patrick E & Joseph L Rodgers. 2018. Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology* 69(1). 487–510.

Suarez-Rivera, Catalina, Jacob L Schatz, Orit Herzberg & Catherine S Tamis-LeMonda. 2022. Joint engagement in the home environment is frequent, multimodal, timely, and structured. *Infancy* 27(2). 232–254.

Suarez-Rivera, Catalina, Linda B Smith & Chen Yu. 2019. Multimodal parent behaviors within joint attention support sustained attention in infants. *Developmental Psychology* 55(1). 96.

Sumer, Omer, Peter Gerjets, Ulrich Trautwein & Enkelejda Kasneci. 2020. Attention flow: End-to-end joint attention estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3327–3336.

Tafasca, Samy, Anshul Gupta & Jean-Marc Odobez. 2023. ChildPlay: A New Benchmark for Understanding Children's Gaze Behaviour. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20935–20946.

Tafasca, Samy, Anshul Gupta & Jean-Marc Odobez. 2024. Sharingan: A Transformer Architecture for Multi-Person Gaze Following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2008–2017.

Tamis-LeMonda, Catherine S, Yana Kuchirko, Rufan Luo, Kelly Escobar & Marc H Bornstein. 2017. Power in methods: Language to infants in structured and naturalistic contexts. *Developmental Science* 20(6). e12456.

Taverna, Andrea, Migdalia Padilla & Sandra Waxman. 2024. How pervasive is joint attention? Mother-child dyads from a Wichi community reveal a different form of "togetherness". *Developmental Science* e13471.

Tomasello, Michael & Michael Jeffrey Farrar. 1986. Joint attention and early language. *Child Development* 1454–1463.

Tomasello, Michael & Jody Todd. 1983. Joint attention and lexical acquisition style. *First language* 4(12). 197–211.

Trevarthen, Colwyn & Penelope Hubley. 1978. Secondary intersubjectivity. *Action, gesture and symbol: The emergence of language* 183–229.

Yu, Chen, Sumarga H Suanda & Linda B Smith. 2019. Infant sustained attention but not joint attention to objects at 9 months predicts vocabulary at 12 and 15 months. *Developmental Science* 22(1). e12735.

Proceedings of the 49th annual Boston University Conference on Language Development

edited by Aditya Yedetore,
Rebecca Dufie Bonney, and Yuanyuan Zhang

Cascadilla Press Somerville, MA 2025

ISSN 1080-692X
ISBN 978-1-57473-037-1 (2 volume set, paperback)

Ordering information

To order a copy of the proceedings or to place a standing order, contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, sales@cascadilla.com, www.cascadilla.com