Modeling joint attention from egocentric vision

Ryan E. Peters[1], **Andrei Amatuni**[1], **Sara E. Schroer**[1], **Shujon Naha**[2], **David Crandall**[2], **Chen Yu**[1]
{ryan.peters, andreiamatuni, saraschroer, chen.yu}@austin.utexas.edu, {snaha, djcran}@iu.edu
[1] Department of Psychology, 108 E Dean Keeton St, Austin, TX 78712
[2] School of Informatics, Computing, and Engineering, 919 E 10th St, Bloomington, IN 47408

Abstract

Numerous studies in cognitive development have provided converging evidence that Joint Attention (JA) is crucial for children to learn about the world together with their parents. However, a closer look reveals that, in the literature, JA has been operationally defined in different ways. For example, some definitions require explicit signals of "awareness" of being in JA—such as gaze following, while others simply define JA as shared gaze to an object or activity. But what if "awareness" is possible without gaze following? The present study examines egocentric images collected via head-mounted eye-trackers during parent-child toy play. A Convolutional Neural Network model was used to process and learn to classify raw egocentric images as JA vs not JA. We demonstrate individual child and parent egocentric views can be classified as being part of a JA bout at above chance levels. This provides new evidence that an individual can be "aware" they are in JA based solely on the in-the-moment visual information. Moreover, both models trained on child views and those trained on parent views leveraged the visual properties associated with visual object holding to improve classification accuracy—suggesting a critical role for object handling in not only establishing JA, as shown in previous research, but also in inferring the social partner's attentional state during JA.

Keywords: joint attention; computational modeling; eye-tracking and visual attention; parent-child social interaction

Introduction

The coordination of parent-child attention shapes infants' early learning experiences. Moments of Joint Attention (JA), or sharing attention to an object or task, provide opportunities for word learning (Tomasello & Farrar, 1986), as well as "scaffolding" to help infants sustain their attention on objects for extended periods of time (Yu & Smith, 2016). The ability of dyads to enter into and maintain JA has also been linked to numerous developmental outcomes, including later language, cognitive, and self-regulation abilities (e.g., Tomasello & Todd, 1983; Mundy & Newell, 2007). JA has been studied at the "macro-level," with researchers defining attention at the timescale of many seconds or minutes, as well as the "micro-level", at the timescale of milliseconds and seconds. Although the importance of JA has been demonstrated across timescales, there is a disagreement in the field regarding the key mechanisms underlying JA and its impacts on later outcomes.

Macro-level definitions of JA require both the parent and the infant to demonstrate an "awareness" of their social partner's engagement (e.g., Tomasello & Farrar, 1986). This perspective typically focuses on gaze-following as a central organizing mechanism but may also necessitate behaviors that explicitly initiate and respond to JA—such as pointing, speaking, or shifting gaze between an object and the social partner (Gabouer, Oghalai, & Bortfeld, 2018). Such definitions paint JA as a "sophisticated" behavior and miss moments when dyads fail to demonstrate awareness, but still share attention to an object. The importance of awareness in macro-level JA is grounded in theories of early word learning. For infants to learn language, they need to form correct object-label mappings. One piece of that puzzle is knowing what a caregiver is looking at as they speak. However, a growing body of evidence shows that gaze following in parent-toddler dyadic interactions is less prevalent than previously assumed (e.g., Deák et al., 2018; Yu & Smith, 2017a), weakening the foundations of these perspectives and leading to contrasting approaches.

Micro-level definitions of JA study dyadic sensory-motor behaviors at the level of milliseconds and seconds, often using head-mounted cameras or eye trackers (e.g., Yu & Smith, 2016). The operational definition of JA at this level often involves a simple rule: are the parent and infant looking at the same object at the same time? This binary decision is made for every frame of the experiment (often at a rate of 30 frames/sec). Frames that match can then be classified as JA and neither member of the dyad needs to show overt awareness of being in JA. Micro-level studies have shown that not only do parents and infants have markedly different views of the worlds (objects tend to be larger and dominate the infant's field of view; Yu & Smith, 2012), but that they flexibly use different multimodal strategies to engage in JA (Yu & Smith, 2017b). Although multimodal behaviors (such as holding the attended object or talking) are not included in the micro-level definition of JA, coinciding behaviors do have real time effects on JA. Object holding seems to play a mechanistic role in both initiating JA (Yu & Smith, 2017a, 2017b) and extending the duration of attention to objects (Suarez-Rivera, Smith, & Yu, 2019). Furthermore, the amount of time a dyad spends attending to each other's hands is even predictive of how often they are in JA (Yu & Smith, 2013, 2017a). For both parents and infants, hands create a pathway into JA, similar

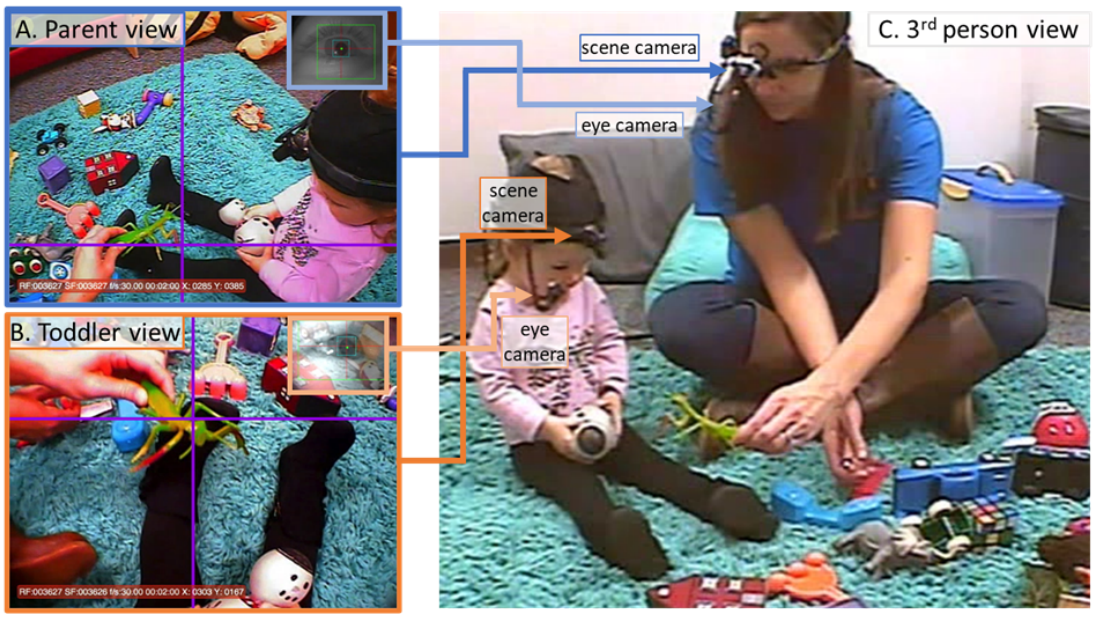

Figure 1: Experimental setup showing (A) first-person parent, (B) first-person toddler and (C) third-person views capturing a moment of Joint Attention (JA) on the praying mantis while the dyad plays together with a set of 24 toys in a naturalistic environment while wearing head-mounted eye-trackers. Crosshairs in the first-person views estimate point of fixation.

to the hypothesized function of gaze-following in macro-level perspectives, but without a need for overt "awareness."

Even at the micro-level, there is evidence showing JA supports infant attention to objects in ways that can predict language outcomes (Yu, Suanda, & Smith, 2019). Thus, building on recent work by Siposowa and Carpenter (2019), we hypothesized that a more graded form of "awareness" must exist—one that does not depend on gaze following. To test this hypothesis, we take a modeling approach that builds on recent advances in computer vision and machine learning. In the field of computer vision, the use of small head-mounted cameras and eye-trackers have allowed for the collection of fine-grained ego-centric views during naturalistic behaviors. Meanwhile, in the field of machine learning, Convolutional Neural Networks (CNNs) have made it possible to use such real-world data in computational models. The combination of egocentric vision and deep learning models offers unique opportunities for cognitive and developmental researchers to analyze egocentric video data collected from young children (e.g., Bambach, Crandall, Smith, & Yu, 2018; Orhan, Gupta, & Lake, 2020; Tsutsui, et al., 2020).

In the current study, we follow this contemporary approach. We first collected a dataset of toddler and parent ego-centric video and gaze data as dyads played with a set of toys in a naturalistic environment (Figure 1). This allowed us to both capture the actual visual experiences of the dyad as they interact and to precisely determine moments of JA at the frame level. We then use this data to build computational models that process the raw sensory data available in each individual's views to classify moments as being in an episode of JA or not. In other words, the first aim is to provide new evidence on whether the visual information perceived from the egocentric view contains signals allowing for infants and/or parents to be "aware" they are in JA with the other during social interactions. Considering the important role of holding objects in establishing JA, our second aim is to investigate whether object holding also provides signals for social partners to infer the attentional state of the other *during* JA, and by doing so, contribute to "awareness" during JA.

Method

Data collection

Twenty-seven toddlers (mean age = 19.22 mos [range: 15.2-24.2]; female = 13) and their parents participated in a study on naturalistic interactions during free toy play (Figure 1). Parent-toddler dyads played on the floor in a naturalistic playroom for an average of 7.51 minutes [range: 3.93-11.64]. Parents were asked to play with a set of 24 toys (initially spread randomly on the floor) like they would at home, but to keep their child sitting on the floor because of the cord attaching the ego-centric camera and eye-tracker to the computer. The egocentric (visual field 108°, 480 x 640 pixels per frame), infrared eye-tracking (Positive Science LLC) and 3rd person cameras all sampled at a rate of 30 Hz. The head-mounted eye-tracking setup and calibration procedure followed validated best practices for achieving the closet approximation to actual ego-centric views and accurate fixation estimates (Slone et al., 2018). After the

experiment, all cameras were synchronized, and software was used to create crosshairs on the parent and toddler egocentric views estimating fixation locations. These generated videos were then used to manually code 25

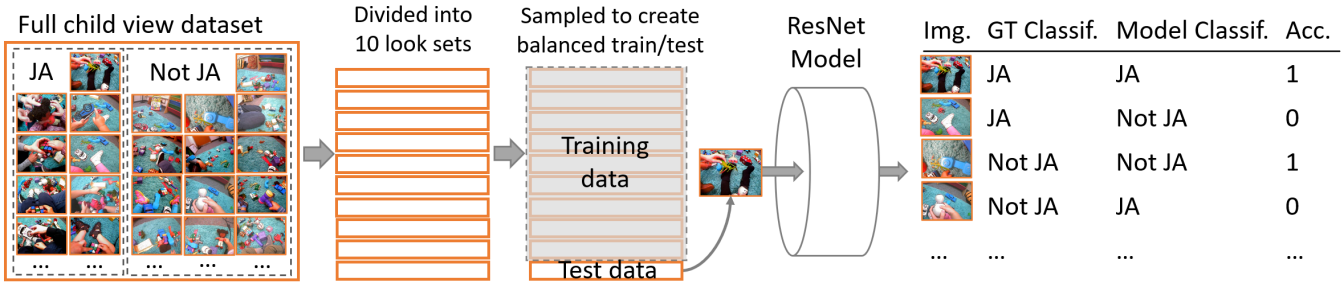

Figure 2: Overview of our modeling pipeline, performed separately for child (shown) and parent models. First, for each subject, "looks" (i.e., sequences of sequential frames with fixations to a particular object) were divided into 10 look sets. Second, balanced look sets were created by sampling 2,000 JA and 2,000 not JA images from each look set. We then used between look training and testing, in which individual images and associated Ground Truth (GT) classifications from 9 out of 10 looks sets were fed into a deep learning (ResNet) model to learn to classify images as JA vs. not JA, after which the model was tested on the remaining look set. Model classifications made on test data were then recorded and used to determine accuracy for each of the 40,000 frames for each dyad member.

regions of interest (the 24 toys and social partner's face), frame by frame, using an in-house coding program.

Detecting Joint Attention

To explore the first aim of determining whether a model trained on raw egocentric toddler or parent views can determine when a frame is actually in JA, hereafter referred to as Ground Truth (GT) JA, bouts of JA were defined as continuous coordination of parent and toddler gaze to the same toy for at least 500ms but including brief looks elsewhere of less 300ms, following precedent (e.g., Yu & Smith, 2017a). However, as the model makes classifications at the level of individual frames, we cleaned the dataset by removing all frames within bouts of JA during which either dyad member was briefly looking away. Furthermore, as we are specifically interested in whether the model can detect JA when the viewer is fixating an object (but not, for example, when they are staring at the wall), we performed the additional cleaning step of removing all frames when the viewer was not fixating an object. This included removing all frames between "looks"—defined as sequences of sequential frames with fixations to a particular experimental object, and all looks to the partner's face. This resulted in a child dataset containing a total of 106,202 JA frames and 152,587 not JA frames, and a parent dataset containing a total of 106,197 JA frames and 113,370 not JA frames.

Computational modeling

As shown in Figure 2, for each dyad member, for each subject, "looks" were randomly assigned to one of 10 look sets.[1] From each look set we then sampled 2,000 JA and 2,000 non-JA images to create 10 balanced look sets (for a total of 40,000 images in the train/test datasets for each dyad member). We then used between look training and testing, in which individual images and associated Ground Truth (GT) classifications from 9 out of 10 looks sets were fed into a deep learning model to learn to classify images as JA vs. not JA, after which the model was tested on the remaining look set. Thus, for each dyad member, 10 models were trained and tested. Also, to be absolutely clear, child models were trained and tested *only* on child egocentric views. Likewise, parent models were trained and tested *only* on parent egocentric views. Our models consisted of state-of-the-art ResNet 50 (He, Zhang, Ren & Sun, 2016) CNNs, trained using stochastic gradient descent implemented via the Adam optimizer. The models were pre-trained on ImageNet (Russakovsky et al., 2015) and the backbones were frozen (i.e., the only trainable parameters were in the final linear layer), so that the network can avoid having to re-learn low-level visual filters from the ground up. The networks output a SoftMax probability distribution over the two classification options: JA and not JA. Images were resized to 224x224 pixels, as required for input to ResNet 50, using bilinear interpolation, and training was performed over 30 epochs with a batch size of 128 and learning rate of 0.0005. Number of epochs and learning rate were chosen based on a grid search centered around default values. Model classifications made on the test data were then recorded and used to determine accuracy for each of the 40,000 frames for each dyad member.

Determining visible holding status of gaze targets

To explore the second aim, investigating whether successful models use object holding as a signal for classifying a moment as JA, for each of the 80,000 images we coded whether the object fixated by the viewer was visible in the image and whether viewer and/or partner were holding the object. This coding is the combination of three basic variables: 1) viewer gaze target, 2) viewer/partner object

[1] Note that, by definition, consecutive images with fixations to a given target object were never put in separate look sets.

holding, and 3) automated object detections. First, for each frame, we determined which object was being fixated by the viewer. Second, we manually coded whether the viewer or partner was holding an object in either hand using an in-house software program, using the synchronized 3rd person views (Fig 1C). Third, we used the well-established YOLO

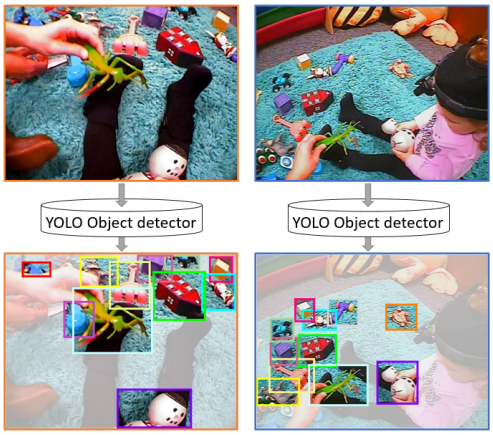

Figure 3: Example of YOLO object detector output for the toddler (left) and parent (right) raw images from figure 1.

object detector (Redmon, Divvala, Girshick, & Farhadi, 2016) to automatically identify objects in each frame and determine whether they were "visible" (Figure 3). For further details, please see the original article for which this was performed (Bambach et al., 2018).

By using the combination of these three variables we were able to pull out four clean visible object holding categories for which we might predict differences in model JA detection performance, and one catch all category for remaining instances that are not of interest for the current research question:

- **Neither holding**: Neither viewer nor partner holding a visible object (i.e., an object detected by the YOLO detector).
- **Only viewer holding**: The viewer is holding and fixating a single visible object, while the partner is not holding a visible object.
- **Only partner holding**: The partner is holding a single visible object that is fixated by the viewer, while the viewer is not holding a visible object.
- **Both holding (the same object)**: The viewer and partner are jointly holding a single visible object that is fixated by the viewer.
- **Other**: All other cases.

Results

Detecting JA from individual egocentric images

Here we address our first aim to examine whether the model can detect whether the social partner was attending to the same object attended by oneself, based solely on in-moment visual information perceived from the egocentric view. To achieve this aim, we first explore whether computational models trained respectively on either toddler or parent egocentric views can learn to classify whether individual images are part of an episode of JA at above chance accuracy. We first calculated the test classification accuracy for each of the 20 ResNet models, and then plotted the means and 95% confidence intervals of those values, by dyad member, in Figure 4.

Two-tailed, one sample t-tests revealed that models trained on both child, MN=0.62, SD=0.02, $t(9)$=16.363, p<.001, d=5.17, and parent data, MN=0.62, SD=0.01, $t(9)$=28.464, p<.001, d=9.0, are able to classify whether egocentric images belong to a bout of JA or not at significantly above chance values. This confirms that there is enough in-the-moment information for toddlers and their parents to be "aware" of the moments they are in JA with their social partner. With such information readily available in their egocentric views, social awareness within JA episodes can be achieved without the need for more sophisticated behaviors such as gaze following.

Furthermore, a Welch two sample t-test comparing mean test classification accuracy for models trained and tested on child images versus those trained and tested on parent images was not significant, $t(14.041)$=0.533, p=.6. In other words, both child and parent views contain visual information that allows for similarly accurate inference of social awareness.

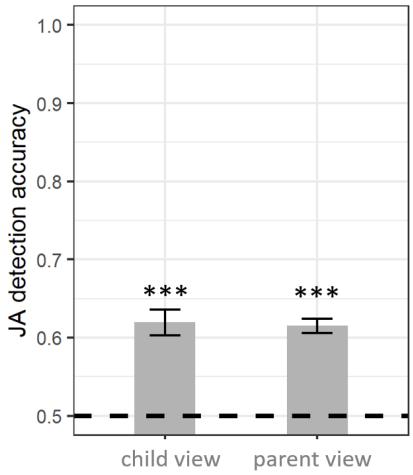

Figure 4: Mean test classification accuracy across the 10 models run for each dyad member. The dashed line shows chance accuracy. Error bars show 95% confidence intervals. *** p < .001.

Examining the role of visible object holding

We next address our second aim of investigating whether object holding provides useful signals for detecting whether the social partner was attending to the same object at the same moment. Previous empirical findings demonstrate that object handling is related to both the initiation of JA bouts

and the extension of child attention to objects. Accordingly, we hypothesize the moments with object handling by either the viewer or partner provide more useful information to detect joint attention than the moments without object handling. To test this hypothesis, we examine whether the accuracy with which models classify images as JA vs not JA is impacted by whether viewer and/or partner are holding visible objects. However, for our models to actually learn to

Table 1: Proportions (normalized by row) and total numbers of child and parent egocentric images coded with each of the four combinations of viewer and partner visible object holding of interest: neither holding, only viewer holding, only partner holding, and both holding; and "other" remaining combinations not of interest to the current work.

	neither	only viewer	only partner	both	other	N
Child viewpoint						
not JA	0.22	0.16	0.09	0.02	0.50	20000
JA	0.06	0.28	0.22	0.08	0.35	20000
N	5635	8933	6267	2027	17138	
Parent viewpoint						
not JA	0.15	0.14	0.07	0.01	0.63	20000
JA	0.04	0.19	0.19	0.06	0.52	20000
N	3870	6450	5263	1422	22992	

use these signals, they must be aligned with the classification task within our training data. Thus, before exploring the impact of visible object holding on classification accuracy, we first characterize the proportions of child and parent images containing the different combinations of viewer and partner visible object holding.

Table 1 shows the proportions and total numbers of child and parent egocentric images coded with each of the four visible object holding categories of interest, and for the catch all "other" category (included for completeness, but not discussed nor included in analyses). We first note that the overall patterns of proportions and total numbers are very similar for child and parent viewpoints. Next, looking at each of the holding combinations, we see that frames with no holding are heavily skewed towards GT not JA images. In contrast, the remaining combinations that include viewer and/or partner holding are all skewed towards GT JA images. In other words, for both child and parent egocentric images, the patterns of visible object holding are aligned with the classification task and could theoretically be used by the ResNet models to determine JA classification.

We next explore whether our models actually make use of visible object holding by viewer and/or partner, for either child or parent egocentric views. To simplify our analyses and inferences, we limit our analyses to images that were actually part of JA bouts (i.e., GT JA images) and that were categorized as one of the four visible object holding categories of interest (i.e., images from the "other" category are not included). We first built a logistic mixed effects model predicting test classification accuracy as a function of the holding variable (with levels: neither holding, only viewer holding, only partner holding, and both holding), dyad member viewpoint, and the interaction of holding and viewpoint variables, with test set as a random effect, using the lme4 package (Bates et al., 2015) in R (R Core Team, 2020). Model fit and pairwise comparisons were estimated using type III ANOVA using Satterthwaite's method via the car (Fox & Weisberg, 2019) and emmeans (Length, 2020) packages. Figure 5 shows the mean test classification accuracies for GT JA images by visible object holding states and dyad viewpoint.

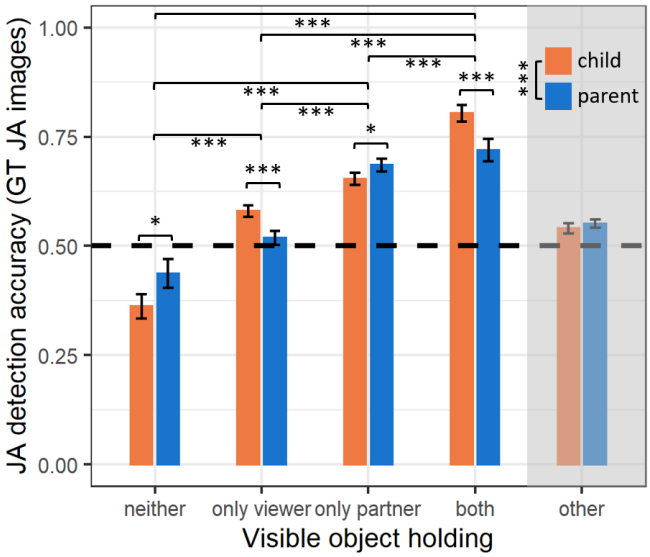

Figure 5: Test detection accuracy for GT JA images by visible object holding state, by dyad member viewpoint, with 95% confidence intervals. The dashed line shows chance accuracy. Accuracy for the "other" holding category is included here for completeness but is not included in analyses.* p < .05. *** p < .001.

Analyses revealed significant main effects of holding state, $\chi^2(3) = 578.17$, $p < .001$, and dyad member viewpoint, $\chi^2(1) = 11.86$, $p < .001$, and a significant interaction, $\chi^2(4) = 76.381$, $p < .001$, on test classification accuracy. Pairwise comparisons across the object holding states revealed the main effect of holding is driven by increased accuracy for viewer, *coef. estimate* = 0.592, *SE* = 0.05, $z = 11.772$, $p < .001$, partner, *coef. estimate* = 1.116, *SE* = 0.0513, $z = 21.754$, $p < .001$, and both holding, *coef. estimate* = 1.574, *SE* = 0.064, $z = 24.758$, $p < .001$, relative to neither holding; greater accuracy for partner, *coef. estimate* = 0.524, *SE* = 0.049, $z = 16.431$, $p < .001$, and both holding, *coef. estimate* = 0.982, *SE* = 0.0492, $z = 19.951$, $p < .001$, relative to only viewer holding; and higher accuracy for both holding relative to partner holding, *coef. estimate* = 0.457, *SE* = 0.05, $z = 9.114$, $p < .001$. Pairwise comparisons of child and parent view within each of the holding states revealed the main effect of dyad viewpoint is driven by greater accuracy for child vs. parent views for viewer holding, *coef. estimate* = 0.242, *SE* = 0.0426, $z = 5.668$, $p < .001$, and both holding, *coef. estimate* = 0.473, *SE* = 0.089, $z = 5.327$, $p < .001$,

though models trained on parent views showed greater accuracy for neither holding, *coef. estimate* = 0.313, *SE* = 0.0909, $z = 3.444$, $p = .013$, and only partner holding, *coef. estimate* = 0.152, *SE* = 0.0471, $z = 3.216$, $p = .028$.

Altogether, these results demonstrate both models trained on child egocentric views and those trained on parent egocentric views make use of visual object holding by the viewer and partner in similar ways to improve classification accuracy for GT JA images. However, models trained on child images are able to make greater use of this information, particularly for images in which only the viewer is holding a visible object and those in which both viewer and partner are jointly holding a visible object.

Discussion

Decades of research has emphasized the importance of "awareness" while engaging in JA. However, recent work has shown that while gaze following in parent-toddler interactions is less prevalent than previously assumed (e.g., Yu & Smith, 2017a), there are still real-time impacts of JA on child behaviors (e.g., Yu & Smith, 2016; Yu et al., 2019). We aimed to use a modeling approach as a first step to examine whether infants and/or parents can detect JA based solely on in-the-moment visual information perceived from egocentric views.

We found that models trained on child or on parent egocentric views were both able to classify whether images were part of a bout of JA. Model classifications were based solely on the visual features directly available in the egocentric views. Our results confirm that it is theoretically possible that dyads can be "aware" of the moments they are in JA without the need for more sophisticated behaviors such as gaze following.

We also present evidence that hand-following is an important cue of JA that is available to children and parents. The child-view and parent-view models both leverage the visual properties associated with visual object holding to improve classification accuracy, as would be expected based on the reviewed empirical results. However, not all visual signals of holding are equally useful—both models trained on child views and those trained on parent views were able to detect JA more accurately when the partner or both the partner and the viewer were holding a visible object. As for the moments when only the viewer was holding an object, only models trained on child views were able to leverage the visual information enough to detect JA at above chance levels. One explanation is that models trained on child views are benefiting from the stronger alignment, relative to parent views, of the only viewer holding category with the modeling task, as shown in table 1. Another non-exclusive explanation is that physical differences between child and parent (e.g., children have shorter arms) could be shaping the visual properties in a way that makes the viewer holding signal clearer in child egocentric images.

These findings have a number of implications for our understanding of JA in naturalistic dyadic interactions. First, these results indicate that coding schemes that actualize JA as a "sophisticated" behavior necessitating explicit signals may be underestimating the amount of time that dyads spend in social joint attention. This inference depends in part on a recently proposed conceptualization of awareness of JA as graded, rather than dichotomous in nature (Siposowa & Carpenter, 2019). This perspective in turn places purportedly qualitatively different types of JA (e.g., triadic vs. shared gaze) on a dynamic continuum that is impacted by correlational patterns with lower-level features immediately available in the visual scene. However, this is not to say that more explicit behaviors are without value. Indeed, such behaviors could be thought of as real-world equivalents of the training signal our models used to learn the associations between JA and the visual features. Crucially, once those associations are learned, being "aware" of being in JA without the need for incessant gaze-following means that children can benefit from the knowledge their parent is looking at the same object while simultaneously engaging in unbroken periods of sustained attention and gaining the consequent positive impacts on learning outcomes (e.g., Yu et al., 2019).

Finally, while much progress has been made in detecting JA from the egocentric views of two or more individuals (e.g., Huang, Cai, & Sato, 2020), it is worth noting that the thus far unexplored task of learning to accurately classify whether an individual is in JA with a social partner solely from their own egocentric views may also lead to interesting applications not only in cognitive science research but in human-computer and human-robot interactions in the real world. As the first attempt at solving this problem, the model we used here is a vanilla implementation of ResNet50 with a frozen, pre-trained backbone and minimal fine-tuning of parameters. Furthermore, it makes no use of cross-frame temporal information, placing a relatively low ceiling on model detection accuracy—due to the relatively shorter timescales of the eye fixations used to define JA compared to the timescales of the behavioral signals available in the egocentric views used by the models to detect JA. Thus, in future we work, we plan on employing more advanced models that allow us to consider visual information spanning multiple frames.

In conclusion, the current work takes a novel computational modeling approach to demonstrate that dyad members could in principle know that their partner is with them without having to explicitly check the other's gaze. In addition to gaze following, there is more than one way to achieve social awareness in joint attention.

Acknowledgements

This work was supported in part by the National Institute of Child Health and Human Development (R01HD074601 and R01HD093792), the National Science Foundation (CAREER IIS-1253549), and the Indiana University Office of the Vice Provost for Research, the College of Arts and Sciences, and the Luddy School of Informatics, Computing, and Engineering through the Emerging Areas of Research Project *Learning: Brains, Machines and Children*.

References

Bambach, S., Crandall, D. J., Smith, L. B. & Yu, C. (2018) Toddler-Inspired Visual Object Learning. *Advances in Neural Information Processing Systems* (NeurIPS), 31.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... & Bolker, M. B. (2015). Package 'lme4'. *Convergence*, 12(1), 2.

Deák, G. O., Krasno, A. M., Jasso, H., & Triesch, J. (2018). What leads to shared attention? Maternal cues and infant responses during object play. *Infancy*, *23*(1), 4-28.

Fox, J., & Weisberg, S. (2019). *An {R} Companion to Applied Regression*. 3rd Edition. Sage publications.

Gabouer, A., Oghalai, J., & Bortfeld, H. (2018). Hearing parents' use of auditory, visual, and tactile cues as a function of child hearing status. *International Journal of Comparative Psychology*, *31*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.

Huang, Y., Cai, M., & Sato, Y. (2020). An Ego-Vision System for Discovering Human Joint Attention. *IEEE Transactions on Human-Machine Systems*, *50*(4), 306-316.

Lenth, R. (2020). *emmeans: estimated marginal means, aka least-squares means*. R package v. 1.4.8.

Mundy, P., & Newell, L. (2007). Attention, joint attention, and social cognition. *Current directions in psychological science*, *16*(5), 269-274.

Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. *Advances in Neural Information Processing Systems*, *33*.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. v. 4.0.2.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779-788.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211-252.

Siposova, B., & Carpenter, M. (2019). A new look at joint attention and common knowledge. *Cognition*, *189*, 260-274.

Slone, L. K., Abney, D. H., Borjon, J. I., Chen, C. H., Franchak, J. M., Pearcy, D., ... & Yu, C. (2018). Gaze in action: Head-mounted eye tracking of children's dynamic visual attention during naturalistic behavior. *Journal of Visualized Experiments,* (141).

Suarez-Rivera, C., Smith, L. B., & Yu, C. (2019). Multimodal parent behaviors within joint attention support sustained attention in infants. *Developmental psychology*, 55(1), 96.

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child development*, 1454-1463.

Tomasello, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First language*, *4*(12), 197-211.

Tsutsui, S., Chandrasekaran, A., Reza, M. A., Crandall, D., & Yu, C. (2020). A Computational Model of Early Word Learning from the Infant's Point of View. In *Annual Conference of the Cognitive Science Society*.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244-262.

Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS one*, *8*(11), e79659.

Yu, C., & Smith, L. B. (2016). The social origins of sustained attention in one-year-old human infants. *Current biology*, *26*(9), 1235-1240.

Yu, C., & Smith, L. B. (2017a). Hand–eye coordination predicts joint attention. *Child development*, *88*(6), 2060-2078.

Yu, C., & Smith, L. B. (2017b). Multiple sensory‐motor pathways lead to coordinated visual attention. *Cognitive science*, *41*, 5-31.

Yu, C., Suanda, S. H., & Smith, L. B. (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.