Discovering the Signatures of Joint Attention in Egocentric Video

Guido Pusiol
guido@cs.stanford.edu
Department of Computer Science
Department of Psychology
Stanford University

Laura Soriano
lsoriano@stanford.edu
Department of Psychology
Stanford University

Li Fei-Fei
feifeili@stanford.edu
Department of Computer Science
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Keywords: Joint attention; computer vision; child development; social cognition.

Introduction

How do young children begin learning the meanings of words? Across cultures, early vocabulary includes names for people, simple social routines, animals, and objects (Tardif et al., 2008), suggesting that the earliest words are learned through interaction and play with others (Bruner, 1985). Identifying a caregiver's intended referent is a critical part of learning meaning within these interactions, and this identification is often accomplished through *joint attention*.

Joint attention describes the situation when both child and caregiver are attending to the same thing and when both know that the other is attending to it (for the remainder of the paper we will talk informally about joint attention—JA—as both the phenomenon and the period of time during which it happens; Carpenter & Liebal, 2011). A typical example of JA is a situation where an adult and child are playing with a toy and the infant alternates gaze between the adult and the toy (Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998).

The capacity for JA gradually develops over the first two years of life and usually begins to emerge between 9 and 12 months of age (Morales et al., 2000), coinciding with the beginnings of language learning. In addition, both the skills that enable JA (e.g. pointing, following a caregiver's gaze to a distal target) and the amount of time that children spend in JA with their caregivers are strong predictors of children's early vocabulary growth (Carpenter et al., 1998; Brooks & Meltzoff, 2008; Tomasello & Todd, 1983).

But how do children *know* that they are in joint attention with a caregiver? From an external perspective, joint attention has typically been defined by a sequence of events: (1) one member of the interaction (child or caregiver) directs the other members attention to an object, (2) both members focus visually on the object, and (3) the child indicates awareness of the caregiver (Tomasello & Farrar, 1986).

Previous work has typically used children's gaze as the main indicator of JA, but, from the perspective of both the child and the data analyst, this method has several issues. First, gaze is neither necessary nor sufficient for JA. It is possible to attend jointly through the hands—as with a child reading a picturebook on a parent's lap—or for the child to follow gaze to a distal target and then signal awareness by moving towards it or reaching for it. Indeed, eye-tracking studies investigating signals to reference find that manual signals are far more effective than gaze in manipulating young children's attention (Yurovsky, Wade, & Frank, 2013). Second, young children may not have perceptual access to their caregiver's gaze most of the time. Recent studies using head-mounted cameras and eye-trackers suggest that children are more often looking at the objects in front of them than at the faces of their caregivers (Smith, Yu, & Pereira, 2011; Franchak, Kretch, Soska, & Adolph, 2011; Frank, Simmons, Yurovsky, & Pusiol, 2013). Third, parents most often look at their children, not at the object they are talking about (Frank, Tenenbaum, & Fernald, 2013). Thus, gaze alone is at best a noisy cue for the identification of JA, either for the child or for the researcher attempting to identify JA in a large dataset.

The goal of our current work is to discover other signals of joint attention. There are two purposes to this investigation. The first is data analytic: A better understanding of how to extract JA episodes from video could be a powerful tool for analyzing video corpora. The second is psychological: The unsupervised extraction of JA episodes from video could give hints regarding robust cues that children might use in addition to, or even in lieu of, gaze.

We use two data sources to gain information about the social interaction between child and caregiver: head-mounted and fixed camera videos. Our approach is unsupervised discovery. We hypothesized that the most effective strategy for capturing JA would be the extraction of high-level, semantic features that correspond relatively closely to the kinds of constructs described in prior work manually coding joint attention (e.g. Tomasello & Todd, 1983). Of course, the challenge is that many such features can be extremely difficult to extract in an automated fashion. To compromise, we identified three features that we could extract with relatively high accuracy in an automated fashion; we hypothesized that each of these might have some relationship to JA: (1) caregivers' faces in the egocentric camera, (2) objects that were in motion due to being actively manipulated, and (3) periods of time during which the child's attention was relatively static.

The plan of the paper is as follows. We begin by describing our dataset, and then we describe how we use computational methods to extract semantic features from these data. We then examine the correlations between these features (and higher-level clusters of these features) and hand-coded joint attentional episodes. Our results suggest that there are a number of redundant perceptual cues to JA, and that some of these may be more readily accessible to children than gaze. In future work, some of these cues could form a robust basis for

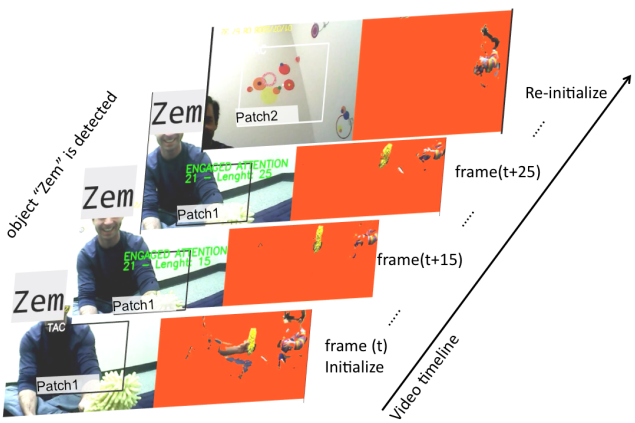


Figure 1: An example of our synchronized dataset: The left side of each panel shows the egocentric video, while the right side shows the motion-filtered 3rd person video. The rectangle in the middle of the egocentric camera shows the attention chunk tracker, while the label "zem" indicates that the object detector has found the yellow feather duster in the 3rd person video.

the automatic detection of joint attentional episodes.

Dataset

We make use of a dataset of in-lab caregiver-child play sessions initially described in Frank, Simmons, et al. (2013). In this dataset, parents were invited to play one-on-one with their children on the floor of a friendly, colorful room. The children wore a small head-mounted (egocentric) camera that captured their approximate visual experience, and a tripod-mounted camera captured the third-person perspective from one corner of the room. Child and caregiver played with a set of toys organized into pairs, with each pair containing a known object (e.g. a ball) and a novel object (e.g. a yellow feather duster). The novel objects were clearly labeled so that parents knew what to call them (e.g. the duster was a "zem"). For purposes of the current study, we chose a set of FIXME videos containing three eight-month-old children and three sixteen-month-old children. Each video was between FIXME and FIXME minutes long.

Annotation of Joint Attention

We used the DataVyu software package (Adolph, Gilmore, Freeman, Sanderson, & Millman, 2012) to annotate periods of time during which child and caregiver were in joint attention. Joint attention was defined if it satisfied the criteria given by (Tomasello & Todd, 1983). First, the interaction had to begin with either parent or child initiating. For example, a parent could hold up an object and label it, or a child could bring an object over the parent. Second, both members were required to focus on the object in JA for at least 3 seconds; we allowed this period to include brief glances away. Third, at some point during the interaction the child was required to display some overt behavior towards the parent to show that he or she acknowledged the interaction.

(a) Motion blur (b) Partial occlusion (c) Low texture

Figure 2: Three examples of challenging faces for traditional detectors.

Defining Semantic Features

We describe automatic and semi-automatic methods for creating high-level semantic features capturing caregivers' faces and episodes of static attention ("attention chunks") from egocentric video and moving objects from the third-person video.

Face Detection

Traditional off-the-shelf face detection algorithms (e.g. Viola & Jones, 2001) fail at detecting parent faces in the kinds of egocentric video that we collected. Face detectors work accurately when the test dataset has low variance from the training dataset and the distance between the camera and the face is >1 meter (e.g. Facebook-style pictures). From the egocentric perspective, however, many other face configurations are prevalent. Faces appear partially occluded or cropped, blurred by motion, and with large size and texture variability making detecting them very challenging (Figure 2).

We addressed the problem using a semi-automated adaptive algorithm (Kalal, Matas, & Mikolajczyk, 2010) that makes use of minimal user input for initialization (selecting one example face per video). The algorithm uses new pixel patches in the trajectory of an optical-flow based tracker to train and update a face detector. The optical flow tracker and the face detector work in parallel. If the face detector finds a location in a new frame exhibiting a high similarity to its stored template, the tracker is re-initialised on that location. Otherwise, the tracker uses the optical flow to decide the location of a face in the new frame.

The primary advantage of the algorithm is the use of motion for face detection: Following the movement of the pixels that define a face it is possible for the algorithm to adapt to new morphologies (i.e. different face poses). More broadly, this method allows for a face that is partially occluded or poorly lit to be tagged as a face by virtue of its relationship with previous frames where the face information was clearer.

Evaluation As part of an ongoing study following Frank, Simmons, et al. (2013), we evaluated this face detector using a set of 37 egocentric videos gathered in the circumstances described above (with ages ranging from 8 – 16 months).

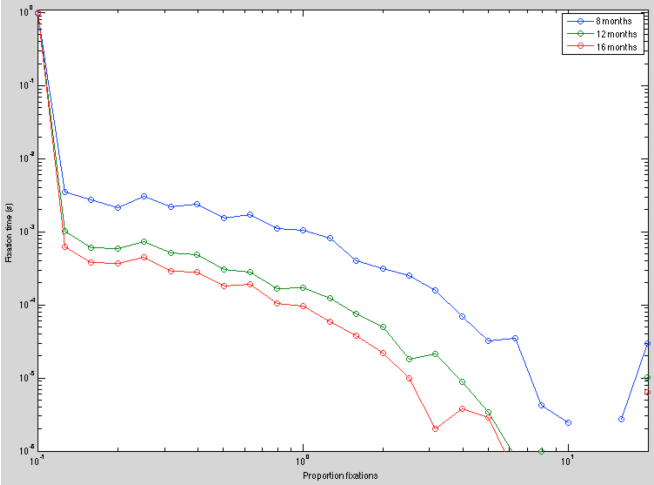


Figure 3: The distribution of attention chunk lengths for 8, 12, and 16 month old children. FIXME

Our evaluation compares automatically detected faces with human ground truth annotations over a sample of both high face-density and randomly selected frames. We found that our algorithm had precision of FIXME and recall of FIXME, achieving a relatively high level of accuracy in this challenging dataset.

Detecting Episodes of Static Attention

One important aspect of joint attention is that it should be (relatively) static if the child is focused on a single object. Congruent with that, previous work has found that episodes where a single object dominates the field of view (and hence the view field is static) are predictive of word learning (Smith et al., 2011; Pereira, Smith, & Yu, 2013). We attempted to identify such moments of fixed attention ("attention chunks") in an automated way. Our strategy is to track a large-scale region of the video (e.g. background texture) across frames; if this texture remains in a relatively static location, we can infer that the child's head has not moved significantly. If the texture deforms substantially, then the head is likely to be in motion. This approach is supported by prior experimental work indicating that eye gaze and head pose are typically coupled (Yoshida & Smith, 2008).

The algorithm is initializes by modeling a pixel-texture patch (P_i). For each new frame the algorithm will seek for a similar patch to the one observed in the previous frame. If the patch is matched, a new point is added to the patch trajectory. If the matching is not achieved, a new patch (P_{i+1}) is learned and the tracking algorithm is re-initialized. The base algorithm used for tracking is a version of a "tracking by detection" algorithm (Kalal et al., 2010). A chunk is defined as the video segment defined by the $start_P$ and end_P frames of the tacked patch trajectory.

Evaluation We evaluated the attention chunk method using the larger dataset of 37 egocentric videos. Average chunk duration is shown in Figure 3. The method yielded a distribution that included many very short chunks (presumably while the head was in motion) as well as some longer episodes of attention. We additionally found that the younger children in the sample (8 months) attended somewhat longer on average. We speculate that this pattern is due to the older children's greater autonomy and mobility.

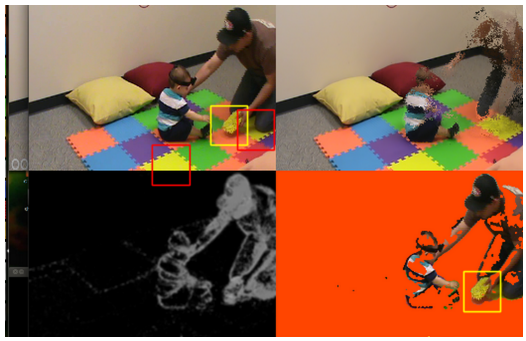
Detection and tracking of moving objects

As described in our earlier work, the vertical field of view of the head-mounted camera is relatively limited ($\sim 40^\circ$ visual angle). Thus, to be able to capture faces high in the visual field, the camera must be at a relatively high angle; this angle in turn precludes capturing the objects that the child is holding. Because of this, we made use of the 3rd person static video to detect the objects that were being handled by the child and the caregiver.

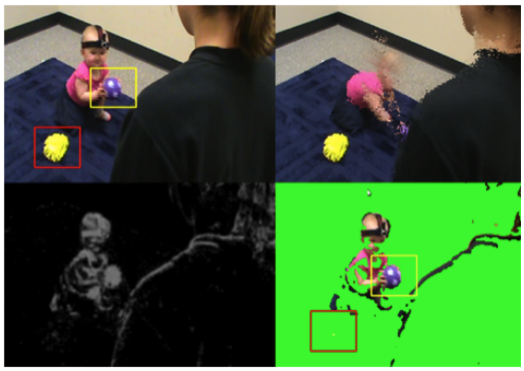
Detection of deformable objects in a colorful, dynamic context is currently an open challenge for computer vision algorithms. Our data contained a wide variety of deformations due to the child-friendly nature of the objects and the consistent occlusion of parts of the objects by caregivers' and children's hands. To circumvent this difficult challenge, we made use of motion as a convenient, psychologically-inspired "filter." Objects that are in motion are more likely to be attended by the child and/or caregiver; in addition, considering only those pixels that are in motion significantly constrains the object-detection problem (Figure 4).

Foreground Modeling The goal of foreground modeling is construct and maintain a statistical representation of the scene so that new information (e.g. due to motion) can be accurately extracted. We chose to utilize both texture information and color information when modeling the background. The approach we use (Yao & Odobez, 2007) exploits the Local Binary Pattern (LBP) feature as a measure of texture because of its good properties (Heikkila & Pietikainen, 2006), along with an illumination invariant photometric distance measure in the RGB space. In brief, this approach computes summary statistics over the static background and searches for local deviations to those summary statistics (due to motion).

Object Tracking We used the extracted foreground pixels as the input to object-tracking algorithms. We experimented with a number of appearance-based object detectors with relatively poor results ("Robust Object Tracking Based on Tracking-Learning-Detection", 2012). Our solution was to detect and track objects by their color and relative size. We modified the cam-shift algorithm (Bradski, 1998), a specialization of the mean-shift algorithm. Mean shift is a non-parametric technique that climbs the gradient of a probability distribution to find the nearest dominant mode (peak). In our case, this distribution is based in color values. The algorithm is initialized by selecting a region containing the object of



(a) Object of interest (marked by the yellow square) could be confused with other objects with similar textures and colors (marked in red). Considering only pixels that are in motion effectively filters these distractors.



(b) Objects of interest are being handled and are therefore moving (yellow square). Unattended objects (red square) are filtered out.

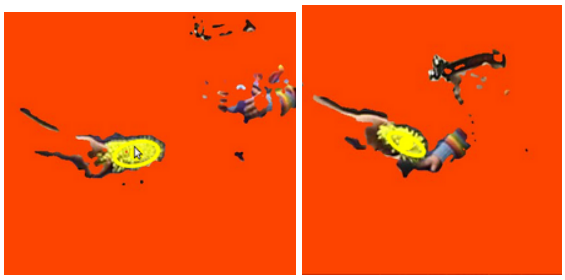
Figure 4: Foreground computation. In each image, top left shows original color frame, top right shows background model color vector, bottom left shows foreground weights, bottom right shows foreground extraction (orange/green pixels correspond to the background mask). (a) shows the problem of texture and color overlap between the object of interest and other objects. (b) shows how this method can also filter out unattended objects.

interest and building a color histogram over the region. In a new frame, the algorithm will match the region's size and the peaks of the color distribution using both mean-shift and euclidean distance metrics. Figure 5 shows examples of the detection and tracking of two different objects.

Feature Aggregation

In the previous sections we detected and computed features (faces, objects, etc.) from different cameras. The goal of the next stage is to merge and prune these features into a single matrix describing the detected features for each video frame. This aggregation required a number of decisions to be made. First, we synchronized frames across the cameras (which had different frame rates). Next, we calculated six features for each frame f (all normalized to the same interval based on the observed maximum and minimum):

(a) "Manu" detected.



(b) "Zem" detected.

Figure 5: Four examples of object detections within the foreground of the static, 3rd person video.

1. Chunk length: The number of frames in the attention chunk containing f_t.
2. Chunk speed: The average speed of the attention chunk's trajectory.
3. Face speed: The speed (L_2 norm) of the face position at f_t and f_{t-1}.
4. Face size: The diagonal of the bounding box for the face (a proxy for distance between parent and child).
5. Object speed: The speed (L_2 norm) of the detected object's position at f_t and f_{t-1}.
6. Object size: The diameter of the ellipse FIXME bounding the detected object.

Because object and face features were computed frame by frame, we experimented with using the attention chunks as a way to propagate features across larger ranges of time. Using this method, all those objects and faces detected in a single video frame that fell within an attention chunk were propagated to all of the frames of the chunk. Assuming that the attention chunks have some value as indicators of the child's attention, this step should improve the quality of detections. We report results both with and without this propagation step.

Evaluation

Independent Feature Analyses

In our first analysis, we examined the proportion of face and object detections that fell inside hand-coded JA episodes. A first indication of the informativeness of these features would be greater proportions of detections within JA. Our findings supported that conclusion: Both faces and objects were more

			Face		Object	
Child	Age	Frames	JA	no JA	JA	no JA
01	8	16783	.08	.04	.11	.11
07	8	19284	.19	.04	.85	.00
15	8	22421	.21	.05	.51	.00
04	16	19337	.29	.16	.44	.00
12-FIXME	16	13450	.08	.04	.11	.11

Table 1: Proportion of detected faces and objects, both inside and outside of JA episodes. Ages are in months.

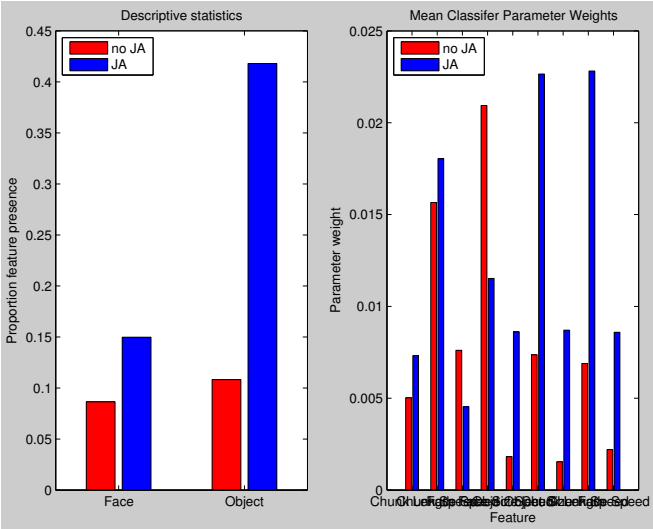


Figure 6: todo...

prevalent within JA episodes for most of the children that we studied (Table 2).

Classification Analysis

Our next analysis used all of the features described above to classify frames as being in or out of joint attention. Although in principle we could have used a complex model that took into account temporal dependencies between frames, we chose to begin by using a simple Naive Bayes classifier. The value of this initial approach is that it allows the straightforward examination of the weights on each feature.

The evaluation of the JA automatic detection is depicted in the confusion matrix. The JA classification is evaluated using cross validation. We evaluate the percentage of correctly classified new frames into JA episodes. The results are 66%, 96%, 94% 95% 60% 88% of correct classification for the 6 test subjects. The 6 subjects can be classified in two age groups, 8 and 16 month old. The average correct classification for the 8 month group is 85 % while for the 16 month group is 81%. The classification drop in the second group occurs due to the children mobility. When children can walk, they can move and play with object without needing the parent engaged in the activity.

		Attn Chunks			Independent		
Child	Age	P	R	F	P	R	F
01	8						
07	8						
15	8						
04	16						
12-FIXME	16						
Total							

Table 2: Precision, recall, and F-score for classifying JA, listed for each child and across all children.

We evaluate the importance (i.e. weight) of each feature learned by the classifier. The figure **??** depicts the weight distribution learned for each independent experiment. In all experiments the features extracted of the detection of objects.

We evaluate the same approach removing the attention chunk shrinking and propagation step. The accuracy of the classifier dropped down an average of 7% in accuracy.

Conclusions

The idea of "joint attention" (JA) is an important construct in understanding children's social interactions with their caregiver. Yet this construct is often defined from the perspective of a knowledgeable third-person observer. Such definitions have both psychological and practical consequences. Psychologically, a growing body of evidence suggests that children may not always have access to their parents' gaze (Franchak et al., 2011; Yu & Smith, 2013; Frank, Simmons, et al., 2013), and so they may have to infer whether they are in joint attention from a host of noisy signals (Frank, Tenenbaum, & Fernald, 2013). Practically, identifying JA in large datasets using automated methods may be exceedingly difficult.

In the current paper, we looked for other features that were related to JA. We computed a set of features that were both semantically meaningful and possible to extract using current methods in computer vision: moments of caregivers' faces, moments of static attention, and objects in motion. We then analyzed the correspondence between these features and hand-coded episodes of JA.

Three findings emerged from this analysis. First, the motion of an object is a simple but highly diagnostic cue to JA—more so than the presence of the caregiver's face. Second, the propagation of features across the "attention chunks" that we identified improved our classification accuracy, suggesting that they carried some information about the child's sustained attention. Third, while the motion of objects was important when children were playing interactively with their parents, older children with greater locomotor ability also played independently, rendering object motion a less reliable cue.

This last trend connects to an important insight from working using egocentric cameras: Children's locomotor development is critically important in determining their view on the social world (Franchak et al., 2011; Frank, Simmons, et al.,

2013; Kretch, Franchak, & Adolph, 2013).

References

Adolph, K. E., Gilmore, R. O., Freeman, C., Sanderson, P., & Millman, D. (2012). Toward open behavioral science. *Psychological Inquiry*, *23*, 244–247.

Bradski, G. R. (1998). Computer Vision Face Tracking For Use in a Perceptual User Interface.

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language*, *35*, 207–220.

Bruner, J. (1985). Child's talk: Learning to use language. *Child Language Teaching and Therapy*, *1*, 111–114.

Carpenter, M., & Liebal, K. (2011). Joint attention, communication, and knowing together in infancy. *Joint Attention: New Developments in Psychology, Philosphy of Mind, and Social Neuroscience*, 159–182.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (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.

Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (2011). Head-mounted eye tracking: A new method to describe infant looking. *Child development*, *82*, 1738–1750.

Frank, M. C., Simmons, K., Yurovsky, D., & Pusiol, G. (2013). Developmental and postural changes in childrens visual access to faces. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society of the.*

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and discourse contributions to the determination of reference in cross-situational word learning. *Language Learning and Development*, *9*, 1–24.

Heikkila, M., & Pietikainen, M. (2006, April). A texture-based method for modeling the background and detecting moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*, 657–662.

Kalal, Z., Matas, J., & Mikolajczyk, K. (2010, June). P-N learning: Bootstrapping binary classifiers by structural constraints. *IEEE Conference on Computer Vision and Pattern Recognition. Proceedings*, 49–56.

Kretch, K. S., Franchak, J. M., & Adolph, K. E. (2013). Crawling and walking infants see the world differently. *Child development*.

Morales, M., Mundy, P., Delgado, C. E., Yale, M., Messinger, D., Neal, R., & Schwartz, H. K. (2000). Responding to joint attention across the 6-through 24-month age period and early language acquisition. *Journal of Applied Developmental Psychology*, *21*, 283–298.

Pereira, A. F., Smith, L. B., & Yu, C. (2013). A bottom-up view of toddler word learning. *Psychonomic bulletin & review*, 1–8.

Robust Object Tracking Based on Tracking-Learning-Detection. (2012, May). , 1–60.

Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mothers view: The dynamics of toddler visual experience. *Developmental science*, *14*, 9–17.

Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N., & Marchman, V. A. (2008). Baby's first 10 words. *Developmental Psychology*, *44*, 929.

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*, 197–211.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer vision and pattern recognition, 2001. cvpr 2001. proceedings of the 2001 ieee computer society conference on* (Vol. 1, pp. I–511).

Yao, J., & Odobez, J. M. (2007). Multi-layer background subtraction based on color and texture. *... and Pattern Recognition*.

Yoshida, H., & Smith, L. B. (2008). What's in view for toddlers? using a head camera to study visual experience. *Infancy*, *13*, 229–248.

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*, e79659.

Yurovsky, D., Wade, A., & Frank, M. C. (2013). Online processing of speech and social information in early word learning. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society of the.*