Discovering the Signatures of Joint Attention in Child-Caregiver Interaction

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

Joint attention—when child and caregiver share attention to an object or location—is an important part of early language learning. Identifying when two people are in joint attention is an important practical question for analyzing large-scale video datasets; in addition, identifying reliable cues to joint attention may provide insights into how children accomplish this feat. We use techniques from computer vision to identify features related to joint attention from both egocentric and fixed-camera videos of children and caregiver interacting with objects. We find that the presence of caregivers’ faces in the child’s egocentric view and the motion of objects in the fixed camera both correlate with human-annotated joint attention. We use a classifier to predict joint attention using these features and find some initial success; in addition, classifier performance is substantially increased by interpolating features across automatically-extracted “attention chunks” in the egocentric video.

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, 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, 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).

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 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 picture-book 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 (Yu & Smith, 2013). 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 (L. B. 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 large 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 & Farrar, 1986). 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: (1) caregivers’ faces in the egocentric camera, (2) objects that were in motion due to being actively manipulated, and (3) periods of time during
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 & Farrar, 1986). 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.

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; Zhu & Ramanan, 2012) perform poorly 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 poorly lit, 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, Mikolajczyk, & Matas, 2012) 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...
Figure 2: Three examples of challenging faces for traditional detectors.

described above (with ages ranging from 8 – 16 months). 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 .90 and recall of .86, 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 (L. B. 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 (W. Smith, 2010). This approach is supported by prior experimental work indicating that eye gaze and head pose are typically coupled (Yoshida & Smith, 2008).

The algorithm is initialized 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 the "tracking by detection" algorithm used above (Kalal et al., 2012). A chunk is defined as the video segment defined by the start and end frames of the tracked patch trajectory.

**Evaluation** We evaluated the attention chunk method using the larger dataset egocentric videos. The distribution of chunk durations 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) had somewhat more long attention chunks; we speculate that this pattern is due to the older children’s greater autonomy and mobility.

![Graph showing the binned distribution of attention chunk lengths for 8-, 12-, and 16-month-old children.](image)

Figure 3: The binned distribution of attention chunk lengths for 8-, 12-, and 16-month-old children. Fixation time and proportion are both plotted on a log scale, because most fixations are very short. The younger (8 month old) children show longer attention fixation episodes compared to the other two groups. A very small number of chunks longer than 10s are not shown.

**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 to construct and maintain a statistical representation of the scene so that new information can be accurately extracted. We chose to utilize both texture information and color information when modeling the background. We follow the approach of Yao and Odobez (2007), which exploits the Local Binary Pattern (LBP) feature as a measure of texture because...
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.

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.

of its good properties (Heikkila & Pietikainen, 2006), along with an illumination invariant photometric distance measure in the RGB space. However, we modified the LBP algorithm to include a larger amount of texture from neighboring pixels. In brief, this approach computes summary statistics over the 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 and experimented with a number of appearance-based object detectors with relatively poor results. Our solution was to detect and track objects by their color and relative size. We modified the camshift 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 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.

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):

1. Chunk length: The number of frames in the attention chunk containing \( f \).
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 \( f \) and \( f_{i-1} \).
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 \( f \) and \( f_{i-1} \).
6. Object size: The maximum diameter of the ellipse containing the pixels of the detected object.
results were shown in Table 1. Although there was substantial variability in accuracy across children, there were no systematic differences across ages. In particular, identifying JA episodes was largely unsuccessful for children like 08-01, while for other children like 08-15 our features were more diagnostic.

We additionally examined the feature weights learned by the classifier. Figure 6B shows average weights on each of the six features. We see very little difference in weight for attention chunk length and speed, suggesting that these characteristics of the chunks did not differ within JA episodes. The most positive weight is given to the object features (especially object speed—which perhaps acted as a proxy for engagement with the object), congruent with the independent feature analyses. Weights for face features mismatched the independent feature analysis, however, with less weight given to faces inside JA episodes. We speculate that this interaction might be due to the fact that faces were present in dyadic play episodes where no object was present as well as in JA episodes.

We evaluate the same approach while removing the attention chunk propagation step. Evaluation metrics for the classifier dropped for nearly all children; in particular, recall dropped considerably (Table 1). While the attention chunks we identified did not directly correlate with JA, they nevertheless provided a useful temporal unit for classification. Further optimizations of the attention chunk detector could improve JA identification.

Conclusions

“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. Two findings emerged from our 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. These findings suggest that, with further development, automated analysis of joint attention from video may be possible in the future.
Figure 6: (A) Proportion of faces and objects detected both in and out of joint attentional episodes. (B) Mean parameter weights on each of the six features we considered for the JA and non-JA categories in the Naive Bayes Classifier. Legend is as in (A).

References


