Discovering new 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 periods of time 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, 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 both strong predictors of children's early vocabulary growth (Carpenter et al., 1998; Brooks & Meltzoff, 2008; Tomasello & Todd, 1983).

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).

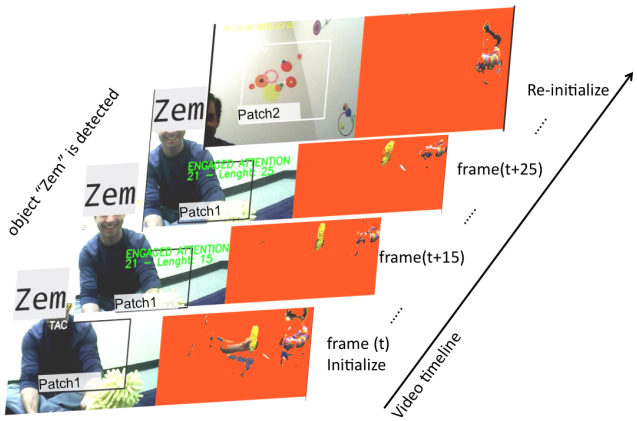
Previous work has typically used children's gaze as the main indicator of JA, but, from the perspective of both the child and the data analyst, this method has several issues. First, gaze is neither necessary nor sufficient for JA. It is possible to attend jointly through the hands—as with a child reading 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 first use computational methods to extract a number of different features—including episodes of static attention, faces, and objects that are currently in motion. 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 robust and readily accessible to children than gaze.

Features Computation

1. Face Detection. Traditional off-the-shelf face detection algorithms fall short to detect the parent faces in the babycam dataset. Face detectors work accurately when the test dataset present low variability on the testing dataset and the distance is above 1 meter (e.g. Facebook like pictures). In the baby perspective other type of face configurations appear. Faces appear partially, blurred and with big size and texture variability (fig. 2) making their detection challenging. We address the problem by with a semi automated adaptive algorithm (Kalal, Matas, & Mikolajczyk, 2010). The algorithm requires manual user input (selecting a face example per



(a) todo

Figure 1: todo...

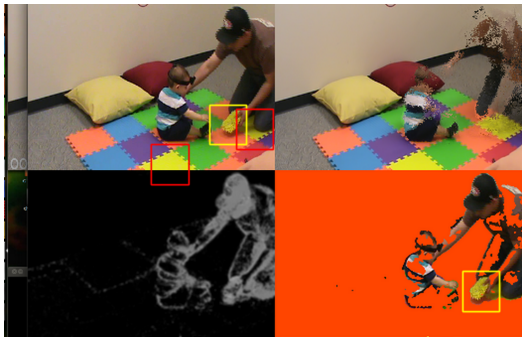
video) for its initialization, but then needed no additional human intervention. The algorithm uses new pixel patches in the trajectory of an optical-flow based tracker to train and update a face detector. The optical flow tracker and the face detector work in parallel. If the face detector finds a location 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). **Precision.** We evaluate the face

(a) Blurred (b) Partial face (c) Low texture

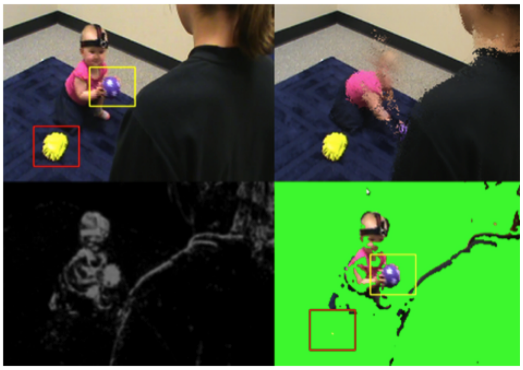
Figure 2: Results (red rectangles) of our face detector in difficult frames.

detector using 37 videos of children's head mounted cameras. Our evaluation compares the automatically detected faces with human ground truth annotations. The metrics: precision = .92; recall = .99; true Negatives = 93; accuracy = 96 outperform the state of the art (Zhu & Ramanan, 2012) due to the active-learning stage of the algorithm.

2. Handled object detection and tracking. We aim at exploring cues of joint attention expressed by hand interactions. The detection of objects being moved even when they are not fully in the visual scope of the children could indicate that the joint attention is established. To automatically track moving objects we first compute the foreground mask representing the moving parts of the video. Understanding the moving parts helps to filter out static patches which can confuse the tracker -figure 6 (a)-, and also to filter out non manipulated objects of interest - figure 6 (b)-



(a) Object of interest (i.e. yellow square) could be confused with other similar textures and color objects in the scene (i.e red squares). The background mask (bottom-right) is filtering the problems out.



(b) Objects of interest are being handled and therefore are moving (i.e. yellow square). Forgotten objects (i.e red squares) are filtered out by the foreground mask.

Figure 3: Foreground computation. In each image, top-left: original color frame, top-right: background model color vector (I), bottom-left: foreground weights, bottom-right: foreground extraction (i.e. orange or green pixels correspond to the background mask). (a) depicts the problem of similar texture and color of the scene and the object of interest. (b) represents the problem of uninteresting objects visible in the scene. Both problems are addressed at once with background/foreground segmentation filtering.

2.1. Foreground Modeling. The goal is to construct and maintain a statistical representation of the scene to be modeled. Here, we chose to utilize both texture information and color information when modeling the background. The approach (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.

The background model $B^t(x)$ of the pixel x at the time t is represented by a list of modes $\{m_k^t(x)\}_{k=1...K}$. The modes register historic information of 7 features of the pixels color and surrounding texture. $m_k^t = \{I, \hat{I}, \check{I}, LBP_k, w, \hat{w}, P\}$. Where I represents the average RGB image vector. $\hat{I}$ and $\check{I}$ are the maximal and minimal RGB image vectors. LBP is the vector of local binary patterns computed at this mode.
The background model can learn up to K^{max} different modes. For each new I^t and LBP^t a new mode is computed m_k^{t-1} and the algorithm seeks to which learned mode m_k^{t-1} the new mode maps to. The mapping is achieved by thresholding a distance (i.e $\tilde{k} = \arg\min_k D(m_k^{t-1}, m_k^t)$). 1) If the new mode cannot be mapped to any of the learned ones and there is still space in the buffer ($K<K^{max}$), then a new mode is initialized. 2) If there is a matched mode $m_{\tilde{k}}^{t-1}$, its representation is updated as follows:

$$\begin{cases} \check{I}_{\tilde{k}}^t = min(I^t, (1+\beta)\check{I}_{\tilde{k}}^{t-1}), \\ \hat{I}_{\tilde{k}}^t = max(I^t, (1-\beta)\hat{I}_{\tilde{k}}^{t-1}), \\ I_{\tilde{k}}^t = (1+\alpha)I_{\tilde{k}}^{t-1} + \alpha I^t, \\ LBP_{\tilde{k}}^t = (1+\alpha)LBP_{\tilde{k}}^{t-1} + \alpha LBP^t, \\ (*)w_{\tilde{k}}^t = (1-\alpha_w^i)w_{\tilde{k}}^{t-1} + \alpha_w^i, \\ \quad \alpha_w^i = \alpha_w(1+\tau\hat{w}_k^{t-1}) \\ \hat{w}_{\tilde{k}}^t = max(\hat{w}_{\tilde{k}}^{t-1}, w_{\tilde{k}}^t) \\ L_{\tilde{k}}^t = 1 + max\{L_k^{t-1}\}_{k=1,...,K,k\neq\tilde{k}}, \\ \quad if \;\; L_{\tilde{k}}^t = 0 \; and \; T_{bw} < \hat{w}_{\tilde{k}}^t \end{cases} \quad (1)$$

where $\beta \in [0,1)$ is the learning rate of the min and max color vectors and $\alpha \in [0,1)$ is the learning rate of the color and texture information. The non matching modes of the previous model are assigned to the new model (i.e. $m^t k := m^{t-1} k$) but their weights are decreased according to (*).
After the update step, all nodes are sorted decreasingly according to their weight. And the background modes are the first B^t modes that satisfy:

$$\sum_{k=1}^{B^t} w_k^t / \sum_{k=1}^{K^t} w_k^t < TB \quad (2)$$

where $TB \in [0,1]$ is the background threshold. Note that the use of both color and texture, the chances that moving foreground objects generate a consistent mode overtime (and beneciate from this effect) are quite small.

2.2 Object tracking. We are interested in detecting a tracking a set of objects (i.e. toys) components of triad when the parents and children are engaged in a JA episode. In particular the toys are being manipulated. The detection and tracking objects is performed over the foreground image. From a computer vision perspective, the objects are highly deformable making them hard to detect and track. The deformations are due to the changes of positions and hand-object occlusions that the object can take while is being manipulated. We have tried different appearance-based ("Robust Object Tracking Based on Tracking-Learning-Detection", 2012) object detectors, and all of them failed. We have finally adopted to detect and track them by its color and relative size. We modified the cam-shift algorithm (Bradski, 1998), which is a specialization of the well known mean-shift algorithm (Comaniciu & Meer, 2002). The mean shift algorithm is a non-parametric technique that climbs the gradient of a probability distribution to find the nearest dominant mode (peak). In our case the distribution is based in color values. The algorithm initializes 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 size and the peaks of the color distribution using mean-shift and the euclidean distance. The figure 4 depicts examples of the detection and tracking of two different objects.

(a) Green object is detected

(b) Zem is detected

Figure 4: (a) the green object is detected, its size and momentum is depicted by the yellow ellipse. (b)

3. Attention Chunks: The attention chunks capture the segments of video of a child engaged visually to a concept. They are extracted of the head camera . The concept can be a concrete object (face, toy, etc.) or an abstract texture captured by the camera when the child's gaze is attending to an off-scope object (e.g. the child is handling a toy near his chest). The ideal case is to have eye trackers to understand the gaze of the baby, but such a configuration is very hard to achieve in uncontrolled environments. Thus, we use the head-cam information as a gaze estimator. Our approach is backed by psychophysical experiments that indicate eye gaze and head pose are coupled in various tasks (Land & Hayhoe, 2001; Pelz & Canosa, 2001). The attention chunk is computed automatically. 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 tracker by detection algorithm (Kalal et al., 2010). An Attention Chunk is the video segment defined by the $start_P$ and end_P frames of the tacked patch trajectory. The figure 5 describes the basics of an attention chunk computation.

We perform analytics over the obtained chunk lengths to understand behavioral structures hidden in the data. We have categorized the extracted chunks by their duration in groups of children of different ages. We used 37 children head can videos (20 min. each) grouped by age: 6, 12 and 16 month old. The average duration of the chunks is depicted in the figure 6(a). Younger children show to have longer episodes of attending to a spot. While the reasons will require of further discussion, our thoughts are that 16 month old children can walk and are more autonomous, and can avoid the parents imposition of attending to certain objects. The figure 6(b) depicts the temporal duration of the chunks distributed in buckets of 10ms in a log scale. The graph shows that the decay is proportionally symmetric among the different groups, most of the attention chunks are of short duration and there are not "distractors" (i.e. a TV) that interfere with our guided attention towards new objects.

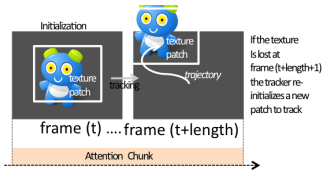
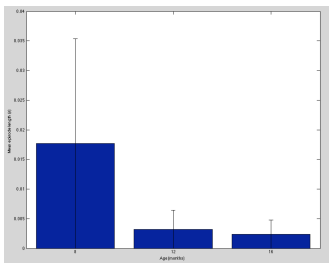
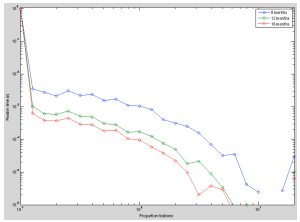


Figure 5: Attention chunk computation.



(a) Duration of attentional chunks ordered by children age 8,12 and 16 month old.



(b) Duration of the attentional chunks distributed in buckets of 10ms and displayed in log scale (i.e. longer chunks appear to the right).

Figure 6: Attention Chunks: analyzing the duration in children of 8,12 and 16 month old.

4. Aggregation: Mapping all features together

In the previous sections we detected and computed features (faces, objects, etc.) in different cameras. The goal of this stage is to merge and prune the features into a single feature matrix describing the detected features at each video frame.

4.1. Synchronization. We calculate the bijective function mapping each frame of the fixed camera to the head mounted camera (i.e. figure). When the inter camera frames can be mapped, the detected features and video attributes of both cameras can also be mapped together.

4.1. Features definition. We calculate 6 features for each video $frame_t$.

- **Chunk Length:** The amount of frames that of the attention chunk containing the $frame_t$.
- **Chunk Speed:** The average speed of the attention chunk trajectory.
- **Face Speed:** The speed, (L_2 norm) of the face position at $frame_t$ and $frame_{t-1}$.
- **Face Size:** The diagonal of the detected face bounding-box. This feature could characterize the parent-child distance.
- **Object Speed:** The speed, (L_2 norm) of the handled object position at $frame_t$ and $frame_{t-1}$.
- **Object Size:** The diameter of the detected object bounding-ellipse.

The lack of detection is replaced by the 0 value.

4.2. Feature shrink an propagation. The object and face features are computed frame by frame. We use the

that are detected in a single video frame occupied by an attention chunk are propagated to all of the frames of the chunk. This step improves the quality of the detection assuming that the attention chunks are strong indicators of the child engaged in a full attention episode.

The operation is achieved bu a composition of mapping functions $f \; o \; g$.

4.3. Normalization. Dealing with multi-modal features requires of a normalization step. We apply the linear mapping of each feature to the range $[0,1)$ using the minimum and maximum values of each feature vector.

Detecting Joint Attention

Is it possible to automatically detect a complex event such as Joint Attention? Being able to extract analytics of big amounts of data, normally impossible for humans to annotate manually can open new research paths in developmental psychology. We show that the descriptive capabilities of the previously computed features is rich enough to train a joint attention classifier and detect these events in new unseen video frames. We train a naive Bayes Classifier. For each children, the training dataset takes half of the positive and negative Joint Attention feature vectors, and we use the remaining half for testing in a cross validation fashion.

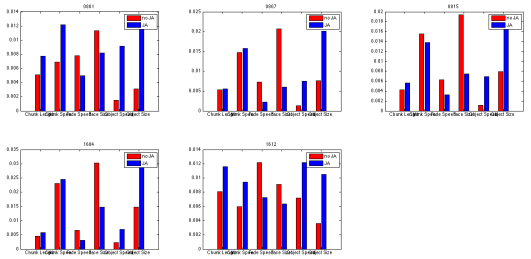
Evaluation

First order analytics can be extracted of the alignment of the automatic feature detection and the annotations. We map into the video timeline the video segments annotated as JA episode and the face and moving object detection. From the alignment, we compute the percentage of time of the automatic detection occurring inside and outside the manually annotated JA episode.

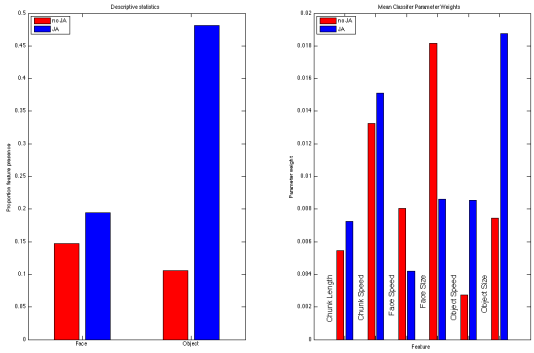
JA Detection Evaluation: The evaluation of the JA automatic detection is depicted in the confusion matrix.

JA Features weights:

The benefits of the attention chunk: We evaluate the same



(a) Weight distribution of the features learned by the model. Each graph represents and an independent experiment, still the inter experiments weight relationship is similar.



(b) Average weights inter models. The results show that the most discriminative feature ha a strong relationship with the handled object, followed by the detection of faces.

Figure 7: todo...

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

Results

Conclusions

This paper aims at better understanding the episodes of joint attention by computing a set of features (i.e. Attention Chunks, Face, Object handling) which can drive to the automatic detection of Joint Attention. In this paper, 3 conclusions emerge. First, the motion of an object that is being part of a child-parent interaction are an important cue for joint attention detection. Second, the detection of sustained attention is a good proxy to improve the detection of joint attention. Third, the motion of objects is important when the children are playing interactively with the parents, but then children have locomotion capabilities the object motion is not a strong cue of joint attention.

References

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.

Comaniciu, D., & Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*, 603–619.

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.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision research*.

Morales, M., Mundy, P., Delgado, C. E., Yale, M., Messinger, D., Neal, R., et al. (2000). Responding to joint attention across the 6-through 24-month age period and early lan-

	First order feature analysis				
	%Faces IN	%Faces OUT	%Toy IN	% Toy OUT	Total frames
Child 0807	19%	4%	85%	$< 0\%$	19284
Child 0815	21%	5%	51%	$< 0\%$	22421
Child 1604	29%	16%	44%	$< 0\%$	19337
Child 0801	8%	4%	11%	11%	16783
Child 1612	8%	4%	11%	11%	13450

Table 1: The table shows the percentage (normalized by time) of detected faces and handled object inside and outside a JA episode. The distribution of faces and handled object is dense inside the JA episode showing that those features are indicators of a JA episode.

guage acquisition. *Journal of Applied Developmental Psychology*, *21*, 283–298.

Pelz, J. B., & Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision research*.

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.

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

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.*

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Cvpr* (pp. 2879–2886). IEEE.

Appendix: Joint Attention presence annotation

Joint attention was defined if it satisfied the criteria by Tomasello and Todd (1983). First, the interaction had to begin with either the parent or child initiating the interaction. 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 focus on the single object for at least 3 seconds (including brief glances away). Third, at some point during the interaction the child must display an overt behavior towards the parent to show that he is aware of the interaction.