# A First-Person Perspective on a Parent-Child Social Interaction During Object Play

**Alfredo F. Pereira, Chen Yu, Linda B. Smith, Hongwei Shen (afpereir@indiana.edu)**
Department of Psychological and Brain Sciences and Department of Cognitive Science
1101 E 10th St, Bloomington, IN 47405 USA

**Abstract**

We studied parent-child social interactions in a naturalistic tabletop setting. Our approach was to analyze the real-time sensorimotor dynamics of the social partners as they engage in joint object play. From the child's point of view, what she perceives critically depends on her own and her social partner's actions, as well as on her actions on objects. These interdependencies may scaffold learning if the perception-action loops, both within a child and between the child and his social partner, can simplify the environment by filtering irrelevant information, for example while learning about objects.

In light of this general hypothesis, we report new findings. These seek to describe the visual learning environment from a young child's point of view and measure the visual information a child perceives in a real-time interaction with a parent. The main results are (1) what the child perceives most often depends on her own actions and her social partner's actions; (2) there are distinct interaction patterns in naturalistic child-parent interactions; these can influence the child's visual perception in different ways; (3) The actions generated by both social partners provide more constrained and clean input to the child, presumably facilitating learning.

**Keywords:** embodied cognition; learning; computational modeling; perception-action.

## Introduction

Consider a small moment in the everyday life of a toddler, one so common as to pass as uneventful: sitting in a living room, playing with toys and with an adult. To the casual observer, watching from the outside as these social partners interact, cognitive development seems quite effortless. Even very young children can engage smoothly in a social exchange, take turns, show bouts of clear joint attention, and switch attention between objects they manipulate and a social partner. Developmental psychologists have long documented that scenes such as this one are not cognitively simple. For example, linking a word to a referent requires the child to succeed at a multitude of tasks: parsing of objects in a scene, motor plans to manipulate objects, establish joint attention, shifts of attention between objects and the social partner, correctly segment and identify the object name from the adult's speech, etc. More importantly, the variability and noise in the scene can be substantial in real world contexts. For example, there could be multiple objects on the floor being seen and talked about, multiple possibilities for action, and no synchrony between the adult's speech and the child's object perception (Baldwin, 1993; Bloom, 2000). Properties of natural interactions such as these, have led many researchers to propose that the child brings into the moment, social, linguistic and conceptual knowledge that constrains attention and learning (Bloom, 2000).

We also view this situation as a context in which there is considerable complexity, but we view the solution to that complexity in terms of the coupled sensory-motor dynamics of the participants. This joint activity happens in the moment and is itself a source of constraints on cognitive development (rather than mere noise to be overcome). Accordingly, our research goal is to examine the natural dynamic structure of real time experience as it evolves in developing children's active engagement with physical objects and with social partners.

## Studying the Sensorimotor Activity of a Parent-Child Dyad

The experimental method we developed, allows researchers to measure, in a laboratory setting, the sensorimotor activity generated by two social partners (Smith, Yu and Pereira, sub.; Yu, Smith, Christensen and Pereira, 2007; Yu, Smith and Pereira, 2008). Our initial goal was to retain the richness of a naturalistic social interaction but exploit the controlled laboratory environment to make large-scale automatic data collection and analysis feasible. The task given to parents was straightforward, play with their child across a small table with toys; a common experience in the lives of children included in this study. Other than sitting at a table and playing with a small set of objects, no more restrictions were placed in the interaction.

### Apparatus: Multimodal Sensing system

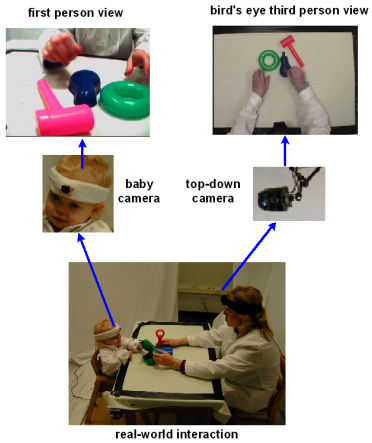


Figure 1: Experimental setup used. The child wears a headband with a small camera and a motion sensor. An additional camera was placed right above the table and provided an unblocked birds-eye view of the interaction.

**Room Setup** A large 6.5m x 3.3m room was split in two. On one side we placed all the recording equipment and an experimenter could sit at a computer to control the experiment. On the other side we created an interaction environment for the parent and child. White, movable curtains, from floor to ceiling, mounted on a ceiling track, surrounded a 3.3m × 3.1m space. At the center of that area we placed a small 61cm × 91cm × 64cm white table. Both participants wore white clothing. A consequence of this setup was that, from the perspective of an head mounted cameras, any pixel not similar to white could only come from an object, a hand or a head of one of the participants.

**Head-Mounted Camera** A small and lightweight camera was mounted on a soft elastic sports headband. The recording rate for the head-mounted camera was 10 frames per second. The resolution of image frame is 320x240. The lens focal length is f3.7mm and the CCD's size is 6.35mm x 6.35mm yielding a field of view of 81 degrees in the horizontal and in the vertical direction. The camera mount piece allowed the camera to rotate on the horizontal axis perpendicular to the camera's line of sight. We used this to adjust the angle between the camera line of sight and the participant's head orientation. Using a similar context of tabletop play, Yoshida and Smith (2008) showed that the head-mounted camera is a good approximation of the contents of the child's visual field. In fact, 90% of head camera video frames in their study corresponded with independently coded eye positions, which was largely due to both the restricted geometry of the tabletop play and the typical behavior of young children in this kind of interaction.

**Bird-Eye View Camera** To provide for an unblocked, static view of the activity on the table we placed a high-resolution camera above it. The high quality of this data stream also meant it was possible to improve the quality of the visual segmentation for the video from a head-mounted camera using information fusion. Also, to calculate what objects were held by each participant.

**Head motion tracking** In addition to the camera mounted in the headband we also mounted a Polhemus motion-tracking sensor to record the child's head position and orientation.

## Data Processing Pipeline: Image Segmentation and Object Detection

The first goal of data processing pipeline was to automatically extract visual information, such as the locations and sizes of objects, hands, and faces, from sensory data in two cameras. These are based on computer vision techniques, and include three major steps (see Figure 2). Given raw images, the first step is to separate background pixels and object pixels. Since we designed the experimental setup (as described above) by covering the walls, the floor and the tabletop with white fabrics and asking participants to wear white cloth, we simply treat close-to-white pixels in an image as background. Occasionally, this approach also removes small portions of an object that have light reflections on them as well (this problem can be fixed in step three). The second step focuses on the remaining non-background pixels and breaks them up into several blobs using a fast and simple segmentation algorithm (Comaniciu and P. Meer, 1997). The third step assigns each blob into an object category. In this object detection task, we used Gaussian mixture models to pre-train a model for each individual object (Moghaddam and Pentland, 1997). As a result of the above steps, we extract useful information from image sequences, such as what objects are in the visual field at each moment, what are the sizes of those objects, and whether a hand is holding an object (from the top-down view), which will be used in the following data analyses. An object is labeled as held by a participant if the object blob is overlapped with a hand blob for more than 10 frames. We asked two human coders to annotate a small proportion of the data (~1200 frames) and compared the results with image processing results with 91% of agreement.

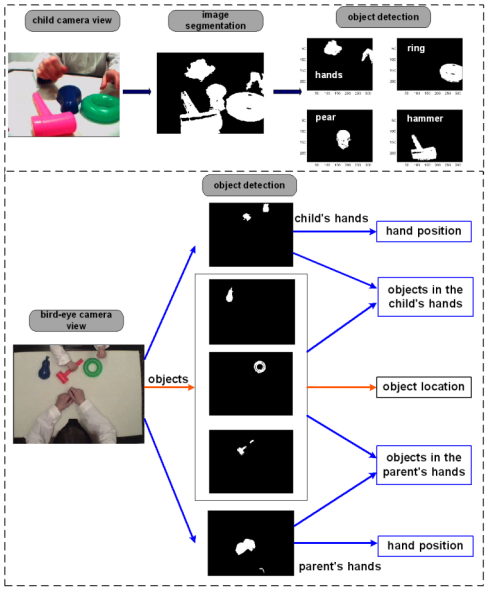


Figure 2: The overview of data processing using computer vision techniques. **Top**: we first remove background pixels from an image and then spot objects and hands in the image based on pre-trained object models. **Bottom**: the processing results from the bird-eye view camera.

## Experimental Study: Object Play Embedded in a Social Interaction With a Mature Partner

**Participants** Fifteen dyads of parent and child participated in this study. Five children were boys. The target age period for this study was 18 to 24 months, a period of large developmental changes in early word learning and visual object recognition (Smith and Pereira, in press). Children's mean age was 21.3 month, ranging from 19.1 to 23.4. All families were from the Bloomington, IN area, white and middle class. Six additional dyads were not included in the

final analysis because the child refused to wear the equipment.

**Stimuli** Parents were given four sets of toys (three toys for each set) in a free-play task. The toys were either rigid plastic objects or plush toys. Most of them had simple shapes and either a single color or an overall main color. Some combinations of objects were selected to elicit actions that were especially evident to an adult asked to play with them.

**Procedure** The procedure needed three experimenters in total. Two of them had specific roles when interacting with parent and child: experimenter (A) focused on providing instructions to the parent and placing the equipment on the child, while another experimenter (B), kept the child distracted continuously until parent and child were ready to start the study. Parents were told in vague terms that the study's goal was to investigate how parent and children interact with the each other. The task was to take three toys from a drawer next to them and play for a fixed period of time. After an indication from the experimenters, they should simply switch to the next set of toys.

When the child seemed well distracted, experimenter (A) placed the headband with the camera and motion sensor in the child's head. This was done quickly and in a single movement that placed camera as close as possible of the child's eyes. Experimenter (B) kept playing and introduced novel toys to keep the child's interest away from the headband.

**Calibration of Head-Mounted Camera** To calibrate the horizontal camera position in the forehead and the angle of the camera relative to the head, the experimenter asked the child to look into one of the objects on the table, placed close to the child. Experimenter (C) controlling the recording in another room confirmed if the object was at the center of the image and if not small adjustments were made on the head-mounted camera gear.

**Trial onset and offset** At this point both experimenters left the room. Parents were told to listen to a specific sound and when alerted by it, start a new trial. This marked trial onset. There were a total of four 90-second trials. The entire study, including initial setup, lasted about 15 minutes.

## Data Analysis and Results: Visual Perception

In this section, we report our results of the child's visual perception, exploring how the child's own actions and the parent's activities may change what the child perceives. We used in total 56 trials from 15 dyads (with 10% of trials from those subjects excluded, due to technical problems such as the quality of data recording). All of the following measures and analyses are trial-based and correspond to averaging sensory data within a trial.

### The Salience of the Visually Dominating Object

We defined the dominating object within a frame as the object that takes the largest proportion in the visual field and is also of a minimum size. We chose a stringent rule: an object is dominant if its ratio inside all object pixels is greater than 0.7 and its total image size is greater than 6000 pixels (7.8% of the image frame). This is a conservative choice of parameters and most likely represents a lower bound on visual salience since this measure just takes in consideration the raw visual size in the image frame.

Our hypothesis was that the child's view might provide a unique window onto the world by filtering irrelevant information (through movement of the body close to the object, or by bringing the object closer), enabling the child to focus on one object (or one event) at a single moment. As shown in Figure 3, in almost 60% of frames, there is one dominating object in the child's view that is much larger than other objects (even when using a conservative rule for what counts as dominating). This result suggests that the child's view may provide a constrained and cleaner input, therefore facilitating learning processes by removing the need to internally handle and filter irrelevant data. If one object (or event) dominates at time, then the focus of learning at the moment is externally decided by the child's bodily selection.

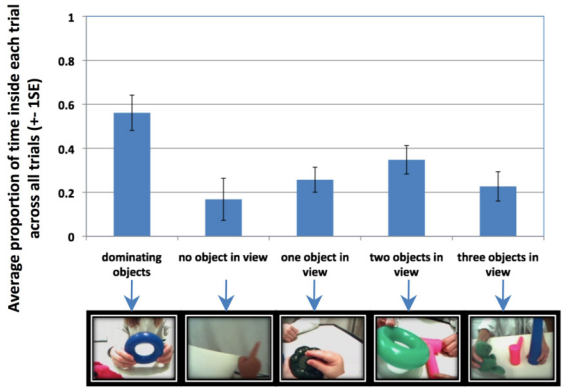


Figure 3: Average proportion of time inside each trial, across all trials where: there is a dominating object; there are no object in the child's view; there are 1,2 or 3 objects in the child's view (the last four categories sum to one). Note that although there are always three objects on the table, in only less than 20% of time, all of the three objects are in the child's visual field while most often there are only 1 or 2 objects (more than 55% in total) in their visual field. Further, in more than 55% of time, there is always a dominating object in the child's visual field at a moment.

### Changes in Dominating Object

The dominating object may change from moment to moment, and also the locations, appearances and the sizes of other objects in the visual field change as well. Thus, we first calculated the number of times that the dominating object changed. From the child's viewpoint, there are on average 15 such object switches in a single trial, suggesting that young children often move their head and body and in

so doing switch attended objects, attending at each moment to just one object.

In summary, the main result so far is that the child's view is far more selective than the available information in the environment (the third-person static view) with respect to (1) the spatial distributions of objects; (2) the salience of individual objects; and (3) the temporal dynamics of the dominating objects.

## Data Analysis and Results: Hand Actions and Visual Perception

Our next analyses considered factors organizing the child's view. At any moment, gaze and head movements may cause changes in the child's visual field. We have directly measured this using a motion tracking system to record head movements in the same setup as the main experiment. The results from that study show that young children rarely move their head toward (closer to) the objects and therefore head movements wouldn't cause the changes of object sizes; that is changes in the dominating object (Pereira, Smith and Yu, 2008). In addition, gaze shifting in this current experiment setup may change the location of an object but not its size in the child's visual field. These results suggest that selectivity, in this setting, happens not through head movements but through hand movements that may bring objects into view. Both the child's own hand movements and the parent's hand movements are potentially relevant since we already know that both participants' hands are frequently appearing in the child's visual field (Yu et al., 2007). In light of this, we measure how hand actions by both the child and the parent select the objects in view for the child.

### Categorizing Parent-Child Interaction Based on who is Holding Objects

In naturalistic interactions, there is variation in the nature of the exchange from person to person across dyads, and from moment to moment even with the same dyad. Some dyads are more actively engaged with the object play task and with potential to-be-learned objects, and others may be less active. For example, how well toddlers and parents coordinate their actions - the smoothness of their "body-prosody" - is correlated with object name learning (Pereira, Smith and Yu, 2008).

Thus, we sought to quantify the quality of dynamic interaction within and across dyads concentrating on how they use their hands to reach, touch, move and manipulate to-be-learned objects, and then used this to quantify the interaction.

We first calculated six individual measures based on what the child and the parent were holding in each moment (sampled at 10Hz): 1) $c_0$: the child is not holding any object. 2) $c_1$: the child is holding one single object. 3) $c_{>1}$: the child is holding more than one objects; 4) $p_0$: the parent is not holding any object; 5) $p_1$: the parent is holding one single object; 6) $p_{>1}$: the parent is holding more than one objects. We then calculated the proportion of time for each event type ($c_0$, $c_1$, $c_{>1}$ and $p_0$, $p_1$, $p_{>1}$) within each trial. By doing so we represented each trial as a 6-dimensional vector, defined by the proportion of time of number of objects held by the child and the parent. Next, we used a hierarchical agglomerative clustering method (Jain and Dubes, 1988) to group 56 interaction trials based on joint hand-state representations described above.

Four primary interaction patterns were found that cover more than 85% (48/56) of all the trials (i.e. they are included in one of the groupings found by the clustering step). These four interaction patterns can be characterized as: 1) child-lead interaction with high activity wherein the child's hands are actively holding and manipulating objects most of the time; 2) bi-directional interaction wherein both the child and the parent are holding objects; 3) parent-lead interaction in which the parent is frequently holding an object; and 4) child-lead interaction with low activity of the participants; both participants are not very active but the child's hands are holding objects more than the parent does. For example, in the trials categorized as the child-lead interaction with high activity, the child is holding only one object more than 64% of time ($c_1$) while the parent does that only 36% of time ($p_1$). In the bi-directional interaction, both the child and the parent spent a significant amount of time on holding one of the three objects on the table ($c_1$ = 45% and $p_1$ = 40%).

Our following data analysis will focus on how both the child's own actions and the parent's actions may influence the child's visual perception in the context of these different interaction patterns.

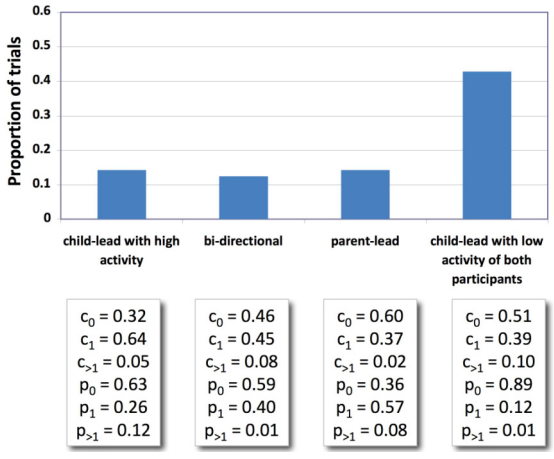


Figure 4: Proportion of trials in each of the four primary interaction patterns identified by the clustering step. A descriptive of each pattern is given below each bar by showing the mean value inside the group for each of the six individual measures ($c_0$, $c_1$, $c_{>1}$ and $p_0$, $p_1$, $p_{>1}$).

### How Often Were the Participants Holding Objects?

As shown in Figure 5 (left), in three out of the four defined interaction patterns, either the child's hands or the parent's hands are holding at least one object more than 70% of time. Critically, hand actions (and other body parts, such as the orientation of the trunk) may signal social cues to the other social partner indicating the object of interest in real time. From this perspective, among these four patterns, the child-lead interaction with high activity and the parent-lead interaction potentially might be seen as creating a better learning environment, wherein the percentage of time that either the child or the parent (at least one of them) is holding an object -- and thus clearly signaling interest in it -- is higher than in the other two interaction patterns.

A relevant result with respect to this idea is that the average percentage of time an object is held in the interaction pattern labeled "child-lead interaction with low activity by both participants" is much lower than in the other three interaction patterns.

As shown in Figure 5 (right panel), the child and parent are holding different objects 28% of time in the parent-lead interaction; this difference in manual selection happens much less frequently in the other three interaction patterns. If hands signal the focus of the attention of the participant, then the child and parent in the parent-lead interaction pattern are often not sharing attention. One possibility is that when the parent attempts to lead the interaction by attracting the child's attention to the object held by the parent, the child does not follow the parent's lead immediately but instead explores other objects that the child himself is interested at the moment. Taken together, our results suggest that hands play an important role in selecting the attended information with the child's own manual activities perhaps being most important. In particular, the child-lead interaction with high activity may provide the cleanest data for the child.

## How Does the Child Perceives Held Objects?

Since both participants' hands are holding objects virtually all the time, we examined how the child perceives an object when the child versus the parent holds it. The first main result is that objects that are held – by either participant – are significantly larger in the child's view (and thus more likely to be dominating the view as defined above).

As shown in Figure 6, objects in hands consistently take a considerable proportion of the child's visual field, calculated as proportion of the field occupied by objects (more than 50%). This is true for all of the four interaction patterns. Even in the child-lead interaction with (relatively) low activity of both participants, objects in hands dominate the visual field in relative terms compared to other objects. The second main result concerns the differences between the child-lead interaction with high activity and the parent-lead interaction: the child's hands play a more important role in influencing the child's own visual field in the first case while the parent's hands are more important in the second case. These results indicate the importance of action – the child's and the parent's – in visual selection, and also raise the question of whether actions by the two participants play different roles for different children, or in different moments of learning.

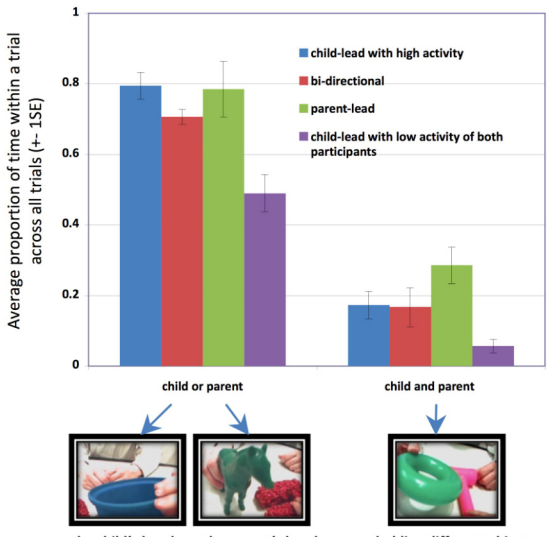


Figure 5: **Left:** the percentage of time that either the child or the parent (at least one of them) is holding an object in each of the interaction pattern. **Right:** the percentage of time that both participants hold objects and that the objects individually held are different, in each of the interaction patterns.

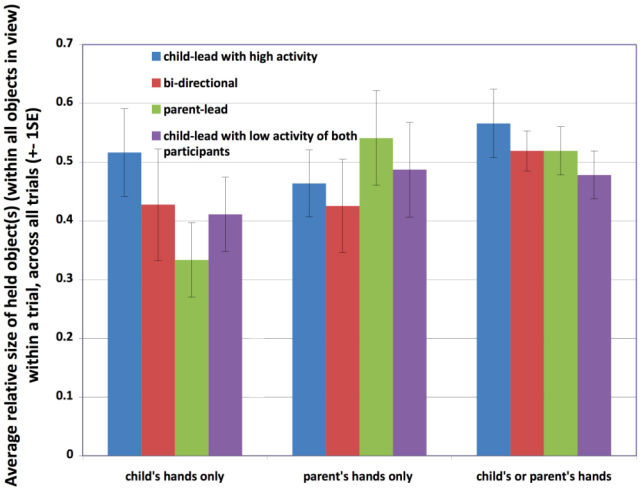


Figure 6: The average size of objects held by the child and/or the parent in the child's visual field (as a proportion of the image frame). There are three conditions: 1) the objects only held by the child; 2) the objects only held by the parent; and 3) the objects held by either the child's or the parent's hands. In all the above cases, we measure the relative proportion of objects in hands compared with other objects in view.

## General Discussion

Viewed from a third-person perspective, the world is a cluttered, noisy, and variable place with many overlapping objects. From this point of view, the learning task seems severe – objects must be segregated one from another, and a speaker's intended referent will be ambiguous. However, the third-person view is not the view that matters for children; the view that matters is their own, the first person view. The visual data collected from a camera placed on a child's forehead suggest that the child's visual field is filtered and narrowed, principally by both the child's own hand actions but also in some cases by the caregiver's actions. Young children solve the problem of noisy and ambiguous data for learning about objects by using the external actions of their own body to select information. This information reduction through their bodily actions most certainly reduces ambiguity and by doing so provide an adaptive strategy to bootstrap learning

The present results also make three points that have not received sufficient attention in the literature. The first concerns the activity of the child in selecting visual information. Considerable research has focused on the social context and how parents select information by guiding the child's attention. But the present results make clear that the child's own activity is also critical and any effect of social scaffolding from the mature partner/ parent will depend on the child's own interests and activities, which may or may not be well coupled to the parents. Second, the results point to manual activities as a major factor in selecting and reducing the visual information. Hands that grab objects and bring them closer to the eyes make those objects large in the visual field and also block the view of other objects, consequences that may benefit many aspects of object recognition (including segregating objects, integrating views, and binding properties) and object name learning. Finally, the results indicate important issues for future work. There may be significant individual differences in how social partners' activities support embodied attention and learning. It could be that some patterns of interaction lead to better learning (e.g., child lead with high activity) than others. Alternatively, it may be that there are different learning styles, each potentially as effective, that engage different mechanisms.

Young children are fast learners and they are so through their embodied interactions with people and objects in a cluttered world. In the present study, we suggest the importance of embodied solutions – how the young child and his social partner may use their bodily actions to *create and dramatically shape* regularities in a learning environment.

## Acknowledgements

This work was supported in part by National Science Foundation Grant BCS0544995, and by NIH grant R21 EY017843. A.F.P. was also supported by the Portuguese Fulbright Commission and the Calouste Gulbenkian Foundation. The authors wish to thank Char Wozniak, Amara Stuehling, Jillian Stansell, Saheun Kim, Heather Elson and Melissa Elson for data collection.

## References

Baldwin, D. (1993). Early referential understanding: Infant's ability to recognize referential acts for what they are. *Developmental psychology*, *29*, 832-843.

Bloom, P. (2000). *How children learn the meanings of words.* Cambridge, MA: The MIT Press.

D. Comaniciu and P. Meer, "Robust analysis of feature spaces: color image segmentation,"Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, pp 750-755, 1997.

Jain, A.K., and Dubes, R.C.,1988, Algorithms for Clustering Data, Prentice Hall: New Jersey.

Moghaddam, B., & Pentland, A. (1997). Probabilistic visual learning for object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(7), 696-710.

Pereira, A. F., Smith, L. B., Yu, C. (2008). Social Coordination in Toddler's Word Learning: Interacting Systems of Perception and Action. Connection Science, 20(2).

Smith, L. B., & Pereira, A. F. (in press). *Shape, action, symbolic play, and words: Overlapping loops of cause and consequence in developmental process*. in S. Johnson (Ed.), A neo-constructivist approach to early development. New York: Oxford University Press.

Smith, L.B., Yu, C. & Pereira, A. F. (under review). Not Your Mother's View: The Dynamics of Toddler Visual Experience. Psychological Science.

Yoshida, H. & Smith, L.B. (2008). Hands in view: Using a head camera to study active vision in toddlers. Infancy.

Yu, C., Smith, L. B., & Pereira, A. (2008). Grounding Word Learning in Multimodal Sensorimotor Interaction. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, Washington, DC, USA.

Yu, C., Smith, L. B., Christensen, M., & Pereira, A. F., (2007). Two Views of the World: Active Vision in Real-World Interaction. In McNamara & Trafton (Eds.), *Proceeding 29nd annual conference of cognitive science society* (p 731-736). Mahwah.NJ: ErIBaum