Two Views of the World: Active Vision in Real-World Interaction

Chen Yu, Linda B. Smith, Mark Christensen, Alfredo Pereira (chenyu@indiana.edu)
Department of Psychological and Brain Sciences, and Cognitive Science Program, Indiana University
Bloomington, IN, 47405 USA

Abstract

An important goal in cognitive development research is an understanding of the real-world physical and social environment in which learning takes place. However, the relevant aspects of this environment for the learner are only those that make contact with the learner's sensory system. In light of this, we report new findings using a novel method that seeks to describe the visual learning environment from a young child's point of view. The method consists of a multi-camera sensing environment consisting of two head-mounted mini cameras that are placed on both the child's and the parent's foreheads respectively. The main result is that the adult's and child's views are fundamentally different in (1) the spatial distributions of hands and everyday objects in the child's visual field and where they are in the parent's field; (2) the salience of individual objects and hands in those two visual fields; and (3) the temporal dynamic structures of objects and hands in two views. These findings have broad implications for how one studies and thinks about developmental processes.

Keywords: Cognitive Development, Embodied Cognition, Joint Attention

Introduction

Children learn about their world – about objects, actions, other social beings, and language -- through their second-by-second, minute-by-minute sensorimotor interactions. Visual information plays a critical role in this early learning. Before babies with normal vision can talk or walk, and before the emergence of any social intelligence to guide their everyday interaction with caregivers, babies are able to perceive and parse their visual environment and are able to move their eyes and head to select visual targets (objects or people) in space. Infants have the opportunity to continuously process complex visual input, and accumulate knowledge from the visual environment. This real time visual information, plus its control through gaze direction and visual attention, contributes to the development of other sensory, cognitive and social capabilities. Indeed, developmentalists such as Gibson (Gibson, 1969) and Ruff (Ruff, 1989) have documented the powerful dynamic visual information that emerges as infants and children move their eyes, heads and bodies, and as they act on objects in the world. In addition, Bertenthal and Campos (1987) have shown how movement – crawling and walking over, under, and around obstacles – creates dynamic visual information crucial to children's developing knowledge about space. Researchers studying the role of social partners in development and problem solving also point to the body and active movement –points, head turns, and eye gaze – in social dynamics and particularly in establishing joint attention (see Smith & Breazeal, 2007, for a review). Computational theorists and roboticists (e.g. Ballard et al., 1997) have also demonstrated the computational advantages of what they call "active vision", how an observer – human or robot – is able to understand a visual environment more effectively and efficiently by interacting with it. This is because perception and action form a closed loop; attentional acts are preparatory to and made manifest in action while also constraining perception in the next moment.

Nonetheless, most previous studies of children's attention and learning have been conducted using macro-level behaviors and in constrained situations, without considering the role of active vision and the perception-action loop. This is in part a consequence of the typical method which uses a third-person camera (or several) to record the child's stream of activities in context. Such recordings provide the view of an outside observer *but not the view of the actively engaged cognitive system.* Further, these views are typically coded by human coders who watch these third person views, a process which is both time consuming and biased, as these coders are *outside observers* with their own psychology and parsing of the events. Understanding how developmental process emerges in second-by-second and minute-by-minute sensorimotor interactions requires capturing (and describing without bias) the first-person view as it is actively generated by the young learner.

The larger goal of this research enterprise is to understand the building blocks for fundamental cognitive capabilities and, in particular, to ground social interaction and the theory of mind in sensorimotor processes. To these ends, we have developed a new method for studying the structure of children's dynamic visual experiences as they relate to children's active participation in a physical and social world. In this paper, we report results from a study that implemented a sensing system for recording the visual input from both the child's point of view and the parent's viewpoint as they engage in toy play. With this new methodology, we compare and analyze the dynamic structure of visual information from these two views. The results show that the dynamic first-person perspective from a child is substantially different from either the parent's or the third-person (experimenter) view commonly used in developmental studies of both the learning environment and parent-child social interaction. The key differences are these: the child's view is much more dynamically variable, more tightly tied to the child's own goal-directed action, and more narrowly focused on the momentary object of interest.

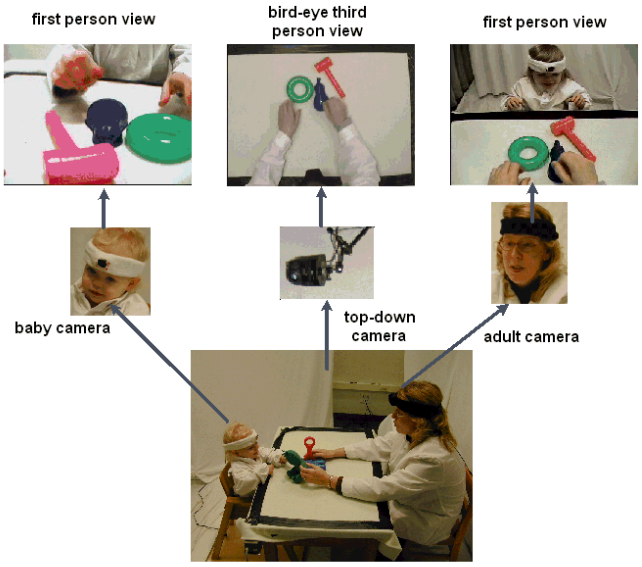

Figure 1: Multi-camera sensing system. The child and the mother play a set of toys at a table. Two mini cameras are placed onto the child's and the mother's heads respectively to collect visual information from two first-person views. A third camera mounted on the top of the table records the bird-eye view of the whole interaction.

Multi-Camera Sensing Environment

The method uses multi-camera sensing system in a laboratory environment wherein children and parents are asked to freely interact with each other. As shown in Figure 1, participant interactions are recorded by three cameras from different perspectives – one head-mounted camera from the child's point of view to obtain an approximation of the child's visual field, one from the parent's viewpoint to obtain an approximation of the parent's visual field, and one from a top-down third-person viewpoint that allows a clear observation of exactly what was on the table at any given moment (mostly the participants' hands and the objects being played with).

Interaction Environment. The study was run in a 3.3m × 3.1m room. At the center of the room a 61cm × 91cm × 64cm table was placed. The table surface was covered in a white soft blanket and the edges were clearly marked with black tape. A high chair for the child and a small chair for the parent was placed facing each other. The walls and floor of the room were covered with white fabrics. Both participants were asked to wear white T-shirts as well. In this way, from any image collected from any camera, white pixels can treated as background while non-white pixels are either objects on the table, the edges of the table, the hands, or the faces of participants.

Head-Mounted Cameras. Two light-weight head-mounted mini cameras (one for the child and another for the parent) were used to record the first-person view from both the child and the parent's perspectives. These cameras were mounted on two everyday sports headbands, each of which was placed on one participant's forehead and close to his eyes. The angle of the camera was adjustable. Input power and video output to these cameras went through a camera cable connected to a wall socket, which was long enough to not cause any movement restriction while participants were sitting down. Both cameras were connected via standard RCA cables to a digital video recorder card in a computer in the room adjacent to the experiment room.

The head camera field is approximately 90 degrees, which is comparable to the visual field of older infants, toddlers and adults (van Hof van Duin & Mohn, 1986; Mohan, Dobson, Harvey, Delaney, & Leber, 1999). One possible concern in the use of a head camera is that the head camera image changes with changes in head movements but not in eye movements. This problem is reduced by the geometry of table-top play. Yoshida & Smith (2007) documented this in a head-camera study of toddlers by independently recording eye-gaze, and showed that small shifts in eye-gaze direction unaccompanied by a head shift do not yield distinct table-top views. Indeed, in their study 90% of head camera video frames corresponded with independently coded eye positions.

Bird-Eye View Camera. A high-resolution camera was mounted right above the table and the table edges aligned with edges of the bird-eye image. This view provided visual information that was independent of gaze and head movements of a participant and therefore it recorded the whole interaction from a third-person static view. An additional benefit of this camera lied in the high-quality video, which made our following image segmentation and object tracking software work more robustly compared with two head-mounted mini cameras. Those two were light-weighted but with a limited resolution and video quality due to the small size.

Parent-Child Joint Interaction Experiment

Participants. The target age period for this study was 18 to 20 months. We invited parents in the Bloomington, Indiana area to participate in the experiment. Nine dyads of parent and child were part of the study. One child was not included because of fussiness before the experiment started. For the child participants included, the mean age was 18.2, ranging from 17.2 to 19.5 months. Three of the included children were female and five were male. All participants were white and middle-class.

Stimuli. Parents were given six sets (three toys for each set) in this free-play task. The toys were either rigid plastic objects or plush toys (three of the total 18). Most of them had simple shapes and either a single color or an overall main color. Some combinations of objects were selected to elicit an action, especially evident to an adult asked to play with them. Figure 2 shows all the stimuli used in this study.

Procedure. The study was conducted by three experimenters: one to distract the child, another to place the head-mounted cameras and a third one to control the quality of video recording. Parents were told that the goal of the study was simply to observe how they interacted with their

Figure 2: Objects used in the parent-child joint interaction study. The ruler shown in each set is 12'' in length.

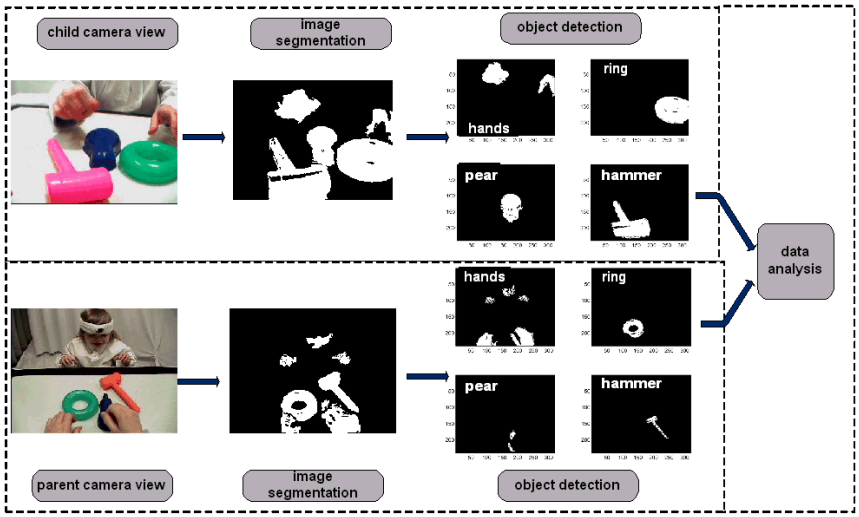

Figure 3: the overview of data processing using computer vision techniques. We first remove background pixels from an image and then spot objects and hands in the image based on pre-trained object models. The visual information from two views is then aligned for further data analyses.

child while playing with toys and that they should try to interact as naturally as possible. Upon entering the experiment room, the child was quickly seated in the high chair and several attractive toys were placed on top of the table. One experimenter played with the child while the second experimenter placed a sports headband with the mini-camera onto the forehead of the child at a moment that he appeared to be well distracted. Our success rate in placing sensors on children is now at over 80%. After this, the second experimenter placed the second head-mounted camera onto the parent's forehead and close to her eyes.

To calibrate the horizontal camera position in the forehead and the angle of the camera relative to the head, the experimenter asked the parent to look into one of the objects on the table, placed close to the child. The third experimenter 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. The same procedure was repeated for the child, with an object close to the child's hands. After this calibration phase, the experimenters removed all objects from the table, asked the parent to start the experiment and left the room. The instructions given to the parent were to take all three objects from one set, place them on the table, play with the child and after hearing a command from the experimenters, remove the objects in this trial and move to the next set to start the next trial. There were a total of six trials, each about 1 minute long. The entire study, including initial setup, lasted for 10 to 15 minutes.

Image Segmentation and Object Detection

The recording rate for each camera is 10 frames per second. In total, we have collected approximately 10800 ($10 \times 60 \times 6 \times 3$) image frames from each interaction. The resolution of image frame is 320×240.

The first goal of data processing is to automatically extract visual information, such as the locations and sizes of objects, hands, and faces, from sensory data in each of three cameras. These are based on computer vision techniques, and include three major steps (see Figure 3). Given raw images from multiple cameras, the first step is to separate background pixels and object pixels. This step is not trivial in general because two first-view cameras attached on the heads of two participants moved around all the time during interaction causing moment-to-moment changes in visual background. However, 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 3). The second step focuses on the remaining non-background pixels and breaks them up into several blobs using a fast and simple segmentation algorithm. This algorithm first creates groups of adjacent pixels that have color values within a small threshold of each other. The algorithm then attempts to create larger groups from the initial groups by using a much tighter threshold. This follow-up step of the algorithm attempts to determine which portions of the image belong to the same object even if that object is broken up visually into multiple segments. For instance, a hand may decompose a single object into several blobs. 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. By applying each object model to a segmented image, a probabilistic map is generated for each object indicating the likelihood of each pixel in an image belongs to this special object. Next, by putting probabilistic maps of all the possible objects together, and by considering spatial coherence of an object, our object detection algorithm assign an object label for each blob in a segmented image as

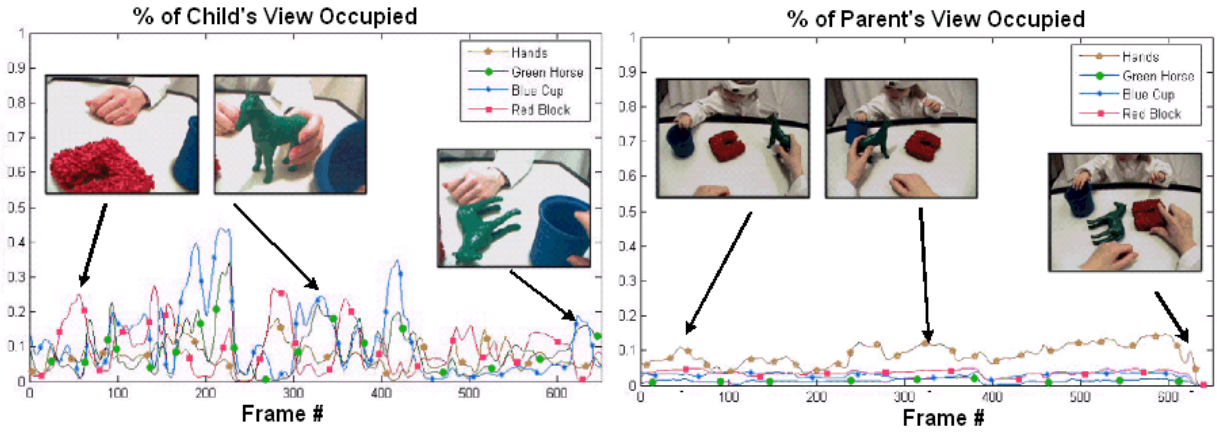

Figure 4: A comparison of the child's and the parent's visual fields. Each curve represents a proportion of an object in the visual field over the whole trial. The total time in a trial is about 1 minute (600 frames). The three snapshots show the image frames from which the visual field information was extracted.

shown in Figure 3. 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, and what are the sizes of those objects, which will be used in the following data analyses.

Data Analyses and Results

The multi-camera sensing environment and computer vision software components enable fine-grained description of child-parent interaction and from two different viewpoints. In this section, we report our preliminary results while focusing on comparing sensory data collected simultaneously from two views. We are particularly interested in the differences between what a child sees and what the mature partner sees.

Visual Field in a First-Person View

The first analyses concern the overall distribution of objects and hands in both the child's and the parent's visual fields: (1) where are hands in the child's visual field and where are they in the parent's field; (2) how those objects occupy the visual fields in two views respectively; and (3) what are the differences of the spatial distributions of objects in those visual fields. Figure 5 shows the overall occupation of objects and hands in two visual fields. The child's view is a closer "zoom-in" view of the physical environment such that objects and hands are all over his visual field. In contrast, there is more structure and regularities in the parent's view. For instance, objects are at the center of her visual field and hands are most often in the peripheral field. Overall, objects and hands are more clustered in the parent's view compared with the one from the child.

A Quantitative Comparison of Two Views

Figure 4 shows the proportion of each object or hand in one's visual field over a whole trial (three snapshots taken from the same moments from these two views). Clearly, the child's visual field is substantially different from the parent's. Objects and hands occupy the majority of the child's visual field and the whole field changes dramatically

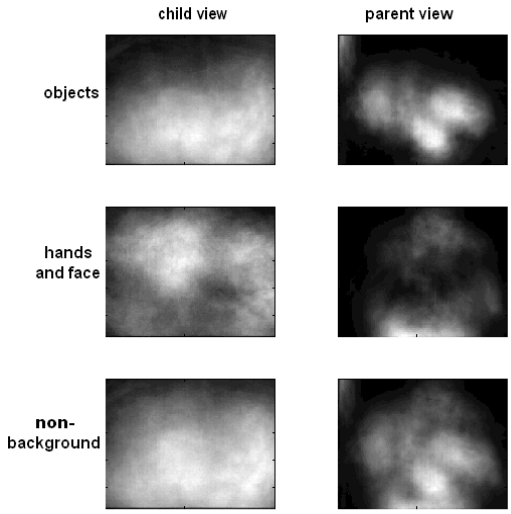

Figure 5: The top row shows the distribution of objects in two first-person views. The second row shows the distribution of hands and faces. The third row is the combination of the two showing the overall distribution of non-background visual information in two visual fields.

moment by moment. In light of this general observation, we developed several metrics to quantify three aspects of the differences between these two views.

First, we measure the composition of visual field shown in Figure 6(a). From the child's perspective, objects occupy about 20% of his visual field. In contrast, they take just less than 10% of the parent's visual field. Although the proportions of hands and faces are similar between these two views, a closer look of data suggests that the mother's face rarely occurs in the child's visual field while the mother's and the child's hands occupy a significant proportion (~15%-35%) in some image frames. From the mother's viewpoint, the child's face is always around the center of the field while the hands of both participants occur frequently but occupy just a small proportion of visual field.

Second, Figure 6(b) compares the salience of the dominating object in two views. The dominating object for a

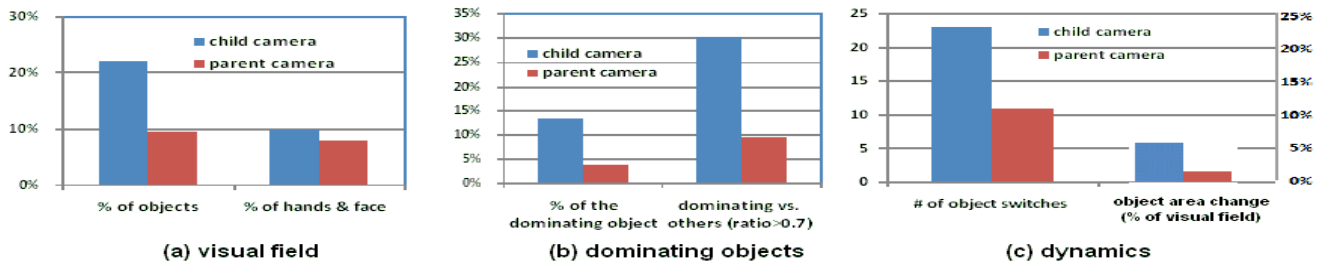

Figure 6: We quantify and compare visual information from two views in three ways. Left: the occupation and composition of visual field. Middle: the salience of dominating objects. Right: the dynamics of visual field over time.

frame is defined as the object that takes the largest proportion of visual field. Our hypothesis is that the child's view may provide a unique window of the world by filtering irrelevant information (through movement of the body close to the object) enabling the child to focus on one object (or one event) at a single moment. To support this argument, the first metric used here is the percentage of the dominating object in the visual field at each moment. In the child's view, the dominating object takes 12% of the visual field on average while it occupies just less than 4% of the parent's field. The second metric measures the ratio of the dominating object vs. other objects in the same visual field, in terms of the occupied proportion in an image frame. A higher ratio would suggest that the dominating object is more salient and distinct among all the objects in the scene. Our results show a big difference between two views. In more than 30% of frames, there is one dominating object in the child's view which is much larger than other objects (ratio > 0.7). In contrast, in less than 10% of time, the same phenomenon happens in the parent's view.

This result suggests not only that children and parents have different views of the environment but also that the child's view may provide more constrained and clean input to facilitate learning processes which don't need to handle a huge amount of irrelevant data because there is just one object (or event) in view at a time. We also note that this phenomena doesn't happen randomly and accidently. Instead, the child most often intentionally moves his body close to the dominating object and/or uses his hands to bring the object closer to his eyes which cause one object to dominate the visual field. Thus, the child's own action has direct influences on his visual perception and most likely also on the underlying learning processes that may be tied to these perception-action loops.

The third measure is the dynamics of visual field, shown in Figure 6(c). The dominating object may change from moment to moment, and also the locations, appearances and the sizes of other objects in the visual field may 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 23 such object switches in a single trial (about 1 minute or 600 frames). There are only 11 per trial from the parent's view. These results together with the measures in Figure 6(b) suggest that children tend to move their head and body frequently to switch attended objects, attending at each moment to just one object. Parents, on the other hand, don't switch attended objects very often and all the objects on the table are in their visual fields almost all the time.

The dynamics of their visual fields in terms of the change of objects in visual field makes the same point. In the child's view, on average, in each frame, 6% of the visual field consists of new objects, objects that are different from the just previous frame to frame. Only less than 2% of the parent's visual field changes this way frame to frame over time. The child's view is more dynamic and that offers potentially more spatio-temporal regularities that may be utilized by leading young learners to pay attention to the more informative (from their point of view) aspects of a cluttered environment.

General Discussion

Embodiment

There are two practical reasons that the child's view is quite different from the parent's view. First, because they are small, their head is close to the tabletop. Therefore, they perceive a "zoom-in", more detailed, and more narrowed view than taller parents. Second, at the behavioral level, children move objects and their own hands close to their eyes while adults rarely do that. Both explanations above can account for dramatic differences between these two views. Both factors highlight the crucial role of the body in human development and learning. The body constrains and narrows visual information perceived by a young learner. One challenge that young children face is the uncertainty and ambiguity inherent to real-world learning contexts: In object recognition, learners need to select the features that are reliably associated with an object from all possible visual features; and in word learning, they need to select the relevant object (at the moment) from among all possible referents on a table. In marked contrast to the mature partner's view, the visual data from the child's first-person view camera suggests a visual field filtered and narrowed by the child's own action. Whereas parents may selectively attend through internal processes that increase and decrease the weights of received sensory information, young children may selectively attend *by using the external actions of their own body*. This information reduction through their bodily actions may remove a certain degree of ambiguity from the child's learning environment and by doing so provide an

advantage to bootstrap learning. This suggests that an adult view of the complexity of learning tasks may often be fundamentally wrong. Young children may not need to deal with all the same complexity from an adult's viewpoint – some of them that complexity may be automatically solved by bodily action and the corresponding sensory constraints.

Joint Interaction

Previous joint-attention research has focused on the temporal synchrony of different participants in real-time interaction. For instance, Butterworth (1991) showed that children and parents share visual attention through social cues signaled by their eyes. Yu, Ballard & Aslin (2005) provided a formal model of the role of gaze in language learning. The present work extends these studies in two important ways. First, our results suggest the importance of spatial information. Children need to not only share visual attention with parents at the right moment; they also need to perceive the right information at the moment. Spatio-temporal synchrony encoded in sensorimotor interaction may provide this. Second, hands (and other body parts, such as the orientation of the body trunk) play a crucial role in signaling social cues to the other social partner. The parent's eyes are rarely in the child's visual field but the parent's and the child's own hands occupy a big proportion of the child's visual field. Moreover, the change of the child's visual field can be caused by gaze and head movement, but this change can also be caused by both his own hand movements and the social partner's hand movements. In these ways, hand movements directly and significantly change the child's view.

A New Window of the World

The first-person view is visual experience as the learner sees it and thus changes with every shift in eye gaze, every head turn, every observed hand action on an object. This view is profoundly different from that of an external observer, the third-person view who watches the learner perform in some environment, precisely because the first person view changes moment-to-moment with the learner's *own* movements. The systematic study of this first person view --- of the dynamic visual world through the developing child's eyes -- seems likely to reveal new insights into the regularities on which learning is based and on the role of action in creating those regularities. The present findings suggest that the visual information from a child's point of view is dramatically different from the parent's (or an experimenter's) viewpoint. This means analyses of third-person views from an adult perspective may be missing the most significant visual information to a young child's learning.

The head camera method used here provides a new look on the structure of the learning environment, and how that structure is generated by the child's own actions. In general, a head camera can provide information about what is in that field --and available to attention -- but does not provide fine-grained information on what the specific focus of the child's attention in that field (as does eye-tracking technology). From this perspective, the head-mounted camera is complimentary to the remote eye-tracking technique which can obtain precise eye gaze location but just in a 2-dimensional pre-defined screen.

Conclusion

The first goal in this work is to see the world as the child sees it and not filtered through our own adult expectations about the structure in that world. The second and equally important goal is to understand how the child's own actions –and coupled actions to a social partner --*create* regularities in visual information. This paper reports beginning progress in reaching these goals and moreover suggests that progress in achieving these goals will bring unexpected new discoveries about the visual environment, about the role of the body, and the structure of the learning task –from the learner's point of view.

Acknowledgments: This research was supported by National Science Foundation Grant BCS0544995 and by NIH grant R21 EY017843.

References

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20* (4), 723-767.

Bertenthal, B. I., Campos, J. J. and Kermoian, R. (1994) An Epigenetic Perspective on the Development of Self-Produced Locomotion and Its Consequences. *Current Directions in Psychological Science, 3* (5) 145-140.

Butterworth (1991) The ontogeny and phylogeny of joint visual attention In *Natural theories of mind: Evolution, development, and simulation of everyday mind reading*, A. Whitten (Eds.), pp. 223- 232, Oxford, England; Blackwell.

Delaney SM, Dobson V, Harvey EM, Mohan KM, Weidenbacher HJ, Leber NR. Stimulus motion increases measured visual field extent in children 3.5 - 30 months of age. *Optom Vis Sci* 2000; 77:82-9.

Gibson, E. J. (1969). Principles of perceptual learning and development. Appleton-Century-Crofts, East Norwalk, CT: US.

Ruff, H.A. (1989). Infants' manipulative exploration of objects: Effects of age and object characteristics. *Developmental Psychology,20*, 9-20.

Smith, L.B. & Breazeal, C. (2007) The dynamic lift of developmental process. *Developmental Science*, 10, 61-68.

van Hof-van Duin & Mohn (1986) The development of visual acuity in normal fullterm and preterm infants. *Vision Research, 26* (6) 909-16.

Yoshida, H. & Smith, L.B. (2007). Hands in view: Using a head camera to study active vision in toddlers. *Infancy.*

Yu, C., Ballard, D.H., & Aslin, R.N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science, 29* (6), 961–1005.