Embodied Solution: The World from a Toddler's Point of View

Chen Yu, Linda B. Smith and Alfredo F. Pereira

Abstract— **An important goal in studying both human intelligence and artificial intelligence is an understanding of how a natural or artificial learning system deals with the uncertainty and ambiguity in the real world. We suggest that the relevant aspects in a learning environment for the learner are only those that make contact with the learner's sensory system. Moreover, in a real-world interaction, what the learner perceives in his sensory system critically depends on both his own and his social partner's actions, and his interactions with the world. In this way, the perception-action loops both within a learner and between the learner and his social partners may provide an embodied solution that significantly simplifies the social and physical learning environment, and filters irrelevant information for a current learning task which ultimately leads to successful learning. 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 results are that (1) the adult's and child's views are fundamentally different when they interact in the same environment; (2) what the child perceives most often depends on his own actions and his social partner's actions; (3) The actions generated by both social partners provide more constrained and clean input to facilitate learning. These findings have broad implications for how one studies and thinks about human and artificial learning systems.**

***Index Terms*— cognitive science, embodied cognition, artificial intelligence.**

I. Introduction

The world's most powerful computers and robots using the most sophisticated software are still far worse than human babies in learning from real world events. Why is this so? One vexing problem for computer scientists is that the real world visual environment is 'cluttered' with lots of overlapping and moving objects, and a computer system simply cannot handle all the information simultaneously available to it. For example, current computer vision systems may be able to learn and recognize several hundreds of 2D objects in supervised mode and in a clean condition (pre-segmented and normalized images) while even 5-year old children can easily recognize thousands of everyday objects in unsupervised mode. Likewise, children seem to learn their native vocabulary with little effort. Meanwhile, there is no existing computational system that can learn and use natural languages in a human-like way (Weng, McClelland, Pentland, Sporns, Stockman, Sur, and Thelen, 2001; Yu, Ballard, and Aslin, 2005; Deak, Barlett, and Jebara, 2007).

What are differences between human learning and machine learning? To deal with noisy data in the real world, most state-of-the-art AI approaches first collect data with (or without) teaching labels from users and the environment, and then rely on implementing advanced mathematical algorithms. These algorithms are then applied to the pre-collected data to induce knowledge. This methodology largely assumes that a learner (e.g. a machine) passively receives information from a teacher (e.g. a human supervisor) in a one-way flow. In contrast, a young child is situated in social contexts and learns about language and about the world through his own actions with the world and with the caregiver. More specifically, the learner actively generates actions to interact with the physical environment, to shape the caregiver's responses, and to acquire just-in-need data for learning. At the same time, the caregiver also dynamically adjusts her behaviours based on her understanding of the learner's state. Thus, the caregiver may provide "on-demand" information for the learner in real time learning. The coupled behaviours between the young learner and the caregiver not only serve as social cues to motivate the learner to be engaged in learning, but also direct the learner's attention to certain aspects in the physical learning environment which will be used as the input to internal cognitive processes (Ballard, Hayhoe, Pook, & Rao, 1997, Triesch, Teuscher, Deak, and Carlson, 2006).

However, most artificial intelligence studies focus on one aspect of learning – what kind of learning device can perform effective computations on pre-collected data, but ignore an equally important aspect of the learning — how a leaner may selectively attend to certain aspects in the learning environment by using his bodily actions to influence his sensory system (the input to the learning device) and also how the caregiver may use her own actions to influence the learner's perception. We suggest here that the relevant aspects in a learning 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 and as well as to measure how the learner's actions and the caregiver's actions may influence the learner's visual information.

This work was supported in part by National Science Foundation Grant BCS0544995 and by NIH grant R21 EY017843.

The three authors are with Department of Psychological and Brain Sciences, Cognitive Science Program, Indiana University, Bloomington, IN, 47405; e-mail: chenyu@indiana.edu).

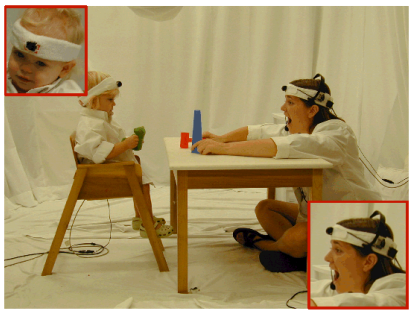

Figure 1: Multi-camera sensing system. The child and the mother play with a set of toys at a table. Two mini cameras are placed onto the child's and the mother's heads respectively to collect visual information from two first-person views. A third camera mounted on the top of the table records the bird-eye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

II. Multi-camera Sensing Environment

The method used a multi-camera sensing system in a laboratory environment wherein children and parents were asked to freely interact with each other. As shown in Figure 1, participants' 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).

A. 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. 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, or the hands, or the faces of participants.

B. 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 to a multi-channel digital video capture card in a recording computer in the room adjacent to the experiment room.

The head camera field is approximately 70 degrees, which is comparable to the visual field of older infants, toddlers and adults. 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. This study 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. Figure 2 (left) shows two snapshots from the two head cameras at the same moment in time.

C. 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. As shown in Figure 2 (right), 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 their small size.

III. 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 a 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.

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 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.

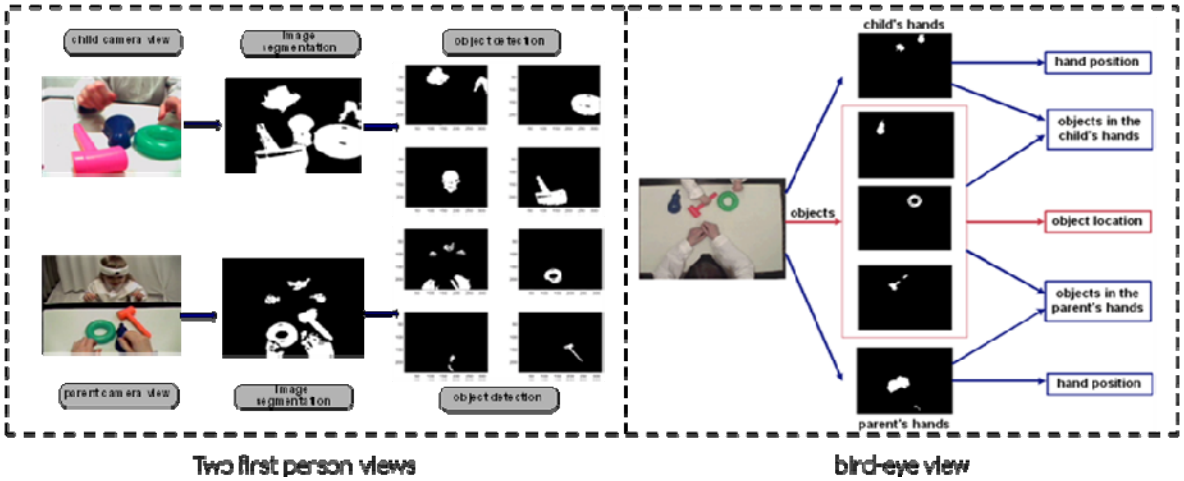

Figure 2: The overview of data processing using computer vision techniques. **Left**: we first remove background pixels from an image and then spot objects and hands in the image based on pre-trained object models. The visual information from two views is then aligned for further data analyses. **Right**: the processing results from the bird-eye view camera.

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.

IV. Image Segmentation and Object Detection

The recording rate for each camera is 10 frames per second. In total, we have collected approximately 10800 (10 × 60 × 6 × 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 shown in Figure 2. 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.

V. Data Analysis 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 results while focusing on comparing sensory data collected simultaneously from two views. We are particularly interested in (1) the differences between what a child sees and what the mature partner sees, and (2) what may cause potential differences (Gibson, 1969).

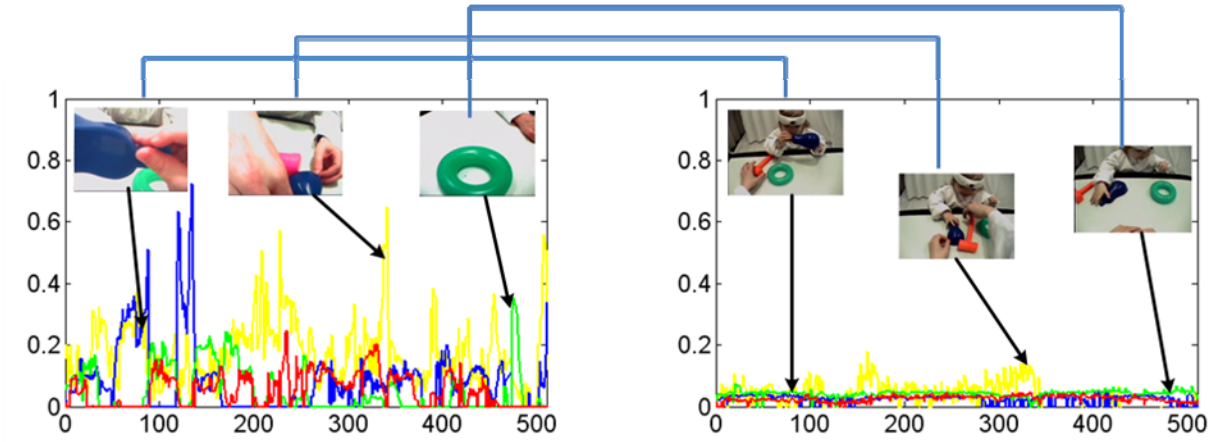

Figure 3: A comparison of the child's and the parent's visual fields. Each curve represents a proportion of an object in the visual field over the whole trial (yellow is hands). The total time in a trial is about 1 minute (500 frames). The three pairs of snapshots from the two views show the image frames from which the visual field information was extracted.

A. Two Different Views of the Same World

Figure 3 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 moment by moment. In light of this general observation, we developed several metrics to quantify three aspects of the differences between these two views.

The composition of visual field. 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. We will further discuss the role of hands in visual perception later.

The salience of the dominating object. The dominating object for a 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 large amount of irrelevant data because there is just one object (or event) in view at a time.

The dynamics of visual field. 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 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.

In summary, the main result so far is that the adult's and child's views are fundamentally different in (1) the spatial distributions of hands and 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 the two views.

B. The Role of the Hands in Visual Perception

What causes the differences between the two views? Why the child's viewpoint is less clustered and more constrained compared with the caregiver's view? There are three possible interpretations. First, the difference may be due to the geometry of the setup. The child is shorter and therefore close to the table while the caregiver is taller and further away from the objects. However, our experimental setup wherein the caregiver and the child sit at the *same* height excludes this possibility. Second, the change of the child's visual field can be caused by gaze and head movement. To test this hypothesis, we have used a motion tracking system to measure head movements in the same setup and the results show that

both participants rarely move their head toward (closer to) the toys. In addition, gaze shifting in this current experiment may change the location of the object but not the size of an object in one's visual field. The third possibility is that the difference may be caused by both the child's own hand movements and the caregiver's hand movements. We already know that both the child's hands and the caregiver's hands are frequently occurring in the child's visual field. More specifically, the parent's eyes are rarely in the child's visual field but meanwhile the parent's and the child's own hands most often occupy a big proportion of the child's visual field. In light of this, we next measure how hands may influence the visual perception of those objects.

What are hands doing? Our first measure shows that both participants actively move their hands and use their hands to grasp and manipulate the objects on the table. As shown in Figure 4, more than 90% of time, at least one object is in either the child's hands or the caregiver's hands. Meanwhile, almost 40% of the time that both the caregiver and the child are holding some objects which can be categorized into two cases: 11% on the same object and 26% on different objects. We suggest that hands (and other body parts, such as the orientation of the trunk) may play several important roles in this toy-play everyday interaction. First, hands may signal social cues to the other social partner indicating the object of interest in real time. Moreover, since their hands are holding objects all the time, how an object is perceived by the child may depend on whether an object is in the child or the caregiver's hands.

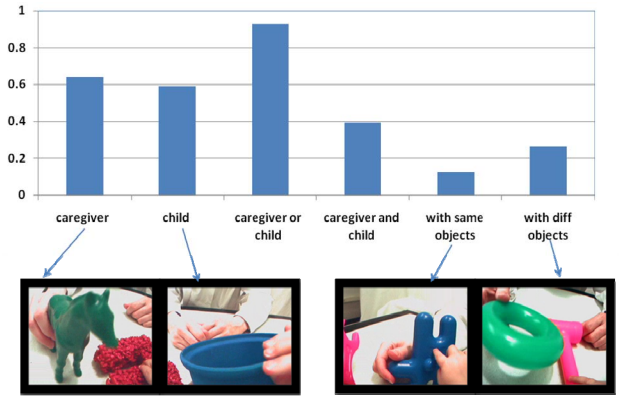

Figure 4: The proportion of time that the child's hands and/or the caregiver's hands are holding objects.

How is an object in hands perceived? Indeed, we find that those objects hold by either the child's hands or the caregiver's hands are significantly larger in the child's view compared with other objects in the environment. As shown in Figure 5, this result further supports our argument that it is *not* because of the geometry differences between the child and the parent that causes the big difference in two views. Instead, the child's own actions and the caregiver's actions determine what the child visually perceives. We also note that this phenomena doesn't happen randomly and accidently. 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; this makes one object dominate the visual field. In addition, to attract the child's attention, the caregiver also moves the object in hands closer to the child's eyes. Thus, both the

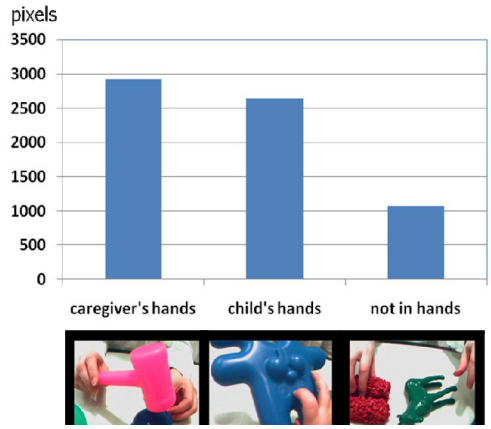

Figure 5: the average sizes of objects in the child's hands, the caregiver's hands or not in anyone's hands from the child's viewpoint.

child's own action and the caregiver's action have direct influences on the child's visual perception and most likely also on the underlying learning processes that may be tied to these perception-action loops.

How does the perception-action loop influence learning? Both artificial and natural learning systems face the uncertainty and ambiguity inherent to real-world learning contexts. In object recognition, a learning system needs to segment a to-be-learned object from the background and extract visual features that are reliably associated with an object category. In word learning, a learning system needs to find the relevant object from all possible referents that are temporally co-occurring with a to-be-learned word. Figure 6 shows that almost all the time that the dominating object (defined above) at a moment from the child's view point is either in the child's hands or the caregiver's hands. Thus, using hands to select one single object in the visual field could facilitate object learning and language learning by allowing learners to focus on one single object at a time. If there is one dominating object at one moment, then human learners can attend to and concentrate on that object – an embodied solution to filter irrelevant information, to disambiguate the

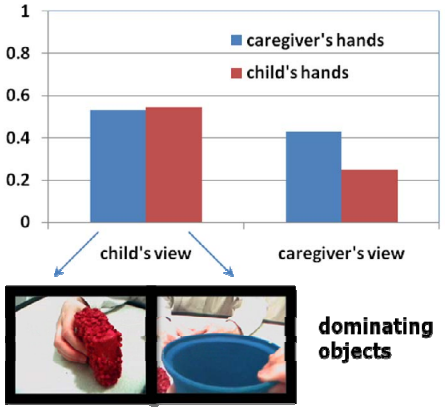

Figure 6: A comparison of the average sizes of objects in the child's hands, in the caregiver's hands or not in hands from the child's viewpoint.

cluttered learning environment, and to learn those objects and names one at a time. Meanwhile, we also note that the caregiver's view point is also influenced by her own actions but not much by the child's actions. Critically, it is because of the child's hands and the caregiver's hands that the dominating object is more salient than other objects in the child's view. The body constrains and narrows visual information perceived by a young learner.

VI. General Discussion and Conclusions

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 both the child's own action and the caregiver's 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 and the bodily actions of the caregiver*. 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. Our result suggests that an adult (e.g. experimenters) 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 that complexity may be automatically solved by bodily action and the corresponding sensory constraints. Hence, the results in the present paper shed lights on a new direction to study powerful human learning – the one based on embodied solution. Here we report beginning progress in reaching these goals and moreover suggest that this new direction will bring unexpected new discoveries about the visual environment from the learner's point of view, about the role of the body, and about the interaction between sensorimotor behaviors and internal learning processes.

A deeper understanding of human learning is directly relevant to building artificial intelligent systems that learn from, teach, and work with humans. Decades of research in artificial intelligence suggest that flexible adaptive systems cannot be fully pre-programmed. Instead, we need to build systems with some preliminary constraints that can create and exploit a rich and variable learning environment. Considerable advances have been made in biologically inspired forms of artificial intelligence (e.g. Breazeal and Scassellati 2002; Asada et al. 2001; Brooks et al. 1998; Steels and Kaplan 2001; Yu, Ballard and Aslin, 2005; Steels and Vogt, 1997; Weng, et al., 2001, Gold and Scassellati, 2007). Young children are fast learners and they do so through their bodily interactions with people and objects in a cluttered world. Could we build a computational system that accomplishes the same learning task? If so, what attributes of a young child are crucial for the machine to emulate? We believe that studies in human learning provide useful hints in various aspects to answer those questions. First, human studies suggest what kinds of technical problems need to be tackled in a computational system. Second, the results from human studies like the present work suggest what are possible solutions employed by human learners and what might be missing mechanisms in the current AI systems. More specifically, we suggest the importance of embodied solution – how the young learner and his social partner may use their bodily actions to *create and dramatically shape* regularities in a learning environment to facilitate learning – which may be a critical component for AI systems to reach human-level intelligence.

Acknowledgment

We thank Amara Stuehling, Jillian Stansell, Saheun Kim, and Mimi Dubner for collection of the data.

References

[1] M. Asada, K. MacDorman, H. Ishiguro and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots", *Robotics and Autonomous Systems*, 37(2), 185-193(9), 2001.

[2] 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.

[3] C. Breazeal, and B. Scassellati, "Infant-like social interactions between a robot and a human caretaker", In *Adaptive Behavior: Simulation Models of Social Agents (special issue)*, 1998.

[4] R.A. Brooks, C. Breazeal, R. Irie, C.C. Kemp and M. Marjanovi. "Alternative essences of intelligence", *In AAAI '98/IAAI '98: Proceedings of the fifteenth national/tenth conference on artificial intelligence/innovative applications of artificial intelligence* (p. 961-968). Menlo Park, CA, USA: American Association for Artificial Intelligence, 1998.

[5] Deák, G.O., Barlett, M. S., and Jebara, T. (2007). "New trends in Cognitive Science: Integrative approaches to learning and development." *Neurocomputing 70*(13-15): 2139-2147 (2007)

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

[7] K. Gold and B. Scassellati (2007). "A Robot that Uses Existing Vocabulary to Infer Non-Visual Word Meanings From Observation." Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI-07), Vancouver, BC, Canada.

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

[9] L. Steels and P. Vogt, Grounding adaptive language game in robotic agents. *In C. Husbands & I. Harvey (Eds.), Proc. of the 4th european conference on artificial life. London: MIT Press*, 1997.

[10] J. Triesch, C. Teuscher, G. Deak and E. Carlson, "Gaze Following: why (not) learn it?", *Developmental Science*, 9(2):125-147, 2006.

[11] Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M. and Thelen, E. (2001), Autonomous mental development by robots and animals. *Science*. v291 I5504. 599-600.

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

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