Published in final edited form as:
IEEE Trans Auton Ment Dev. 2009 August 1; 1(2): 141–151. doi:10.1109/TAMD.2009.2031513.

Active Information Selection: Visual Attention Through the Hands

Chen Yu, **Linda B. Smith**, **Hongwei Shen**, **Alfredo F. Pereira**, and **Thomas Smith**
Psychological and Brain Sciences Department, Cognitive Science Program, Indiana University, Bloomington, IN 47405 USA

Abstract

An important goal in studying both human intelligence and artificial intelligence is to understand how a natural or an artificial learning system deals with the uncertainty and ambiguity of the real world. For a natural intelligence system such as a human toddler, the relevant aspects in a learning environment are only those that make contact with the learner's sensory system. In real-world interactions, what the child perceives critically depends on his own actions as these actions bring information into and out of the learner's sensory field. The present analyses indicate how, in the case of a toddler playing with toys, these perception-action loops may simplify the learning environment by selecting relevant information and filtering irrelevant information. This paper reports new findings using a novel method that seeks to describe the visual learning environment from a young child's point of view and measures the visual information that a child perceives in real-time toy play with a parent. The main results are 1) what the child perceives primarily depends on his own actions but also his social partner's actions; 2) manual actions, in particular, play a critical role in creating visual experiences in which one object dominates; 3) this selecting and filtering of visual objects through the actions of the child provides more constrained and clean input that seems likely to facilitate cognitive learning processes. These findings have broad implications for how one studies and thinks about human and artificial learning systems.

Index Terms

Artificial intelligence; cognitive science; embodied cognition

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. One vexing problem for computer scientists is that the real-world visual environment is 'cluttered' with overlapping and moving objects. Thus, while current computer vision systems can learn and recognize several hundreds of two-dimensional visual objects, they generally require pre-segmented and normalized images; that is, they require cleaned-up input. In contrast, young children seem to easily recognize everyday objects in a cluttered, noisy, dynamic, and three-dimensional world [1]–[6].

One relevant difference between machine learning and human learning that may contribute to this skill gap is the nature of the visual input itself. To deal with noisy data in the real world, most state-of-the-art AI approaches first collect data (with or without teaching labels) and then relies on inventing advanced mathematical algorithms which can be applied to the pre-collected data. The learning system itself is passive in this approach, receiving information in a one-way flow. In contrast, young children learn through their own actions, actions that directly determine what they see, how they see it, and when they see it. Through body movements, young learners actively create the visual input on which object learning depends. If we are to build artificial devices that can—on their own—learn as well as toddlers, we may benefit from understanding

just how young human learners select information through their own actions and how this information selection is accomplished in everyday interactions with adult social partners [7]–[12].

This is also a critical question for theories of human learning. In developmental psychology, many accounts of the learning environment (and the need for constraints on learning mechanisms to learn from that assumed environment, e.g., [13]) are based on adult intuitions about the structure of experience (e.g., [14] in [15]) and not on the actual structure of the child's experiences.

This study examines the toddler's visual environment from the toddler's perspective and asks how toddlers' own actions may play a role in selecting visual information. The data collection and analyses were specifically designed to answer two questions: First, given a cluttered environment, does the child, through their own actions, effectively reduce the information available? Second, what actions by the child are critical to the selection of visual information? In general, four kinds of actions would seem to be relevant to visual selection 1) shifts in eye gaze, 2) head and body turns, 3) manual actions by the child that bring objects into and out of view, and 4) manual actions by the social partner that put objects into the child's view. Here we provide evidence on all of these actions relevant to visual selection, except those due to rapid shifts in eye gaze direction that occur *without* corresponding head turns. We do this because the role of larger body movements in selecting visual information in naturalistic settings has not been studied even though they are likely to be important given the continuous and large body movements characteristic of toddlers. Indeed, recent psychophysics studies on adults performing everyday tasks, such as making a peanut butter and jelly sandwich or making tea, show a close coupling between eye, head, and hand movements with eye gaze slightly leading the head and then the hands [16], [17]. Moreover, recent evidence from developmental studies also suggests that when toddlers are manually acting on objects (not just looking), head and eye shifts are tightly coupled, with most looks to an object also involving simultaneous head turns and posture changes [18]. Therefore, the present study focuses on large body movements, such as head turns and manual actions, and we consider the possible additional contributions of more subtle eye gaze shifts in discussions of the findings.

II. MultiCamera Sensing Environment

To capture the global changes in the information available to the child's visual system as a result of the child's own body, head, and hand actions, we developed a new measurement device that records the available visual information from the child's perspective: as shown in Fig. 1, a mini-camera mounted on a headband. For a stable record of the information available in the environment independent of the child's movements, we used an additional camera placed above the table.

A. The Environment

The study was conducted in a 3.3 m × 3.1 m room. At the center of the room was a 61 cm × 91 cm × 64 cm table painted a nonglossy white. A high chair for the child and a small chair for the parent were placed facing each other. The walls and floor of the room were covered with white fabrics. Both participants were asked to wear white shirts as well. Thus from both cameras, white pixels can be treated as background while nonwhite pixels are either objects on the table, or the hands, or the faces of participants.

B. Head-Mounted Camera

A lightweight head-mounted mini-camera was used to record the first-person view from the young child's perspective which was mounted a sports headband placed on the participant's

forehead and close to her eyes. The angle of the camera was adjustable. Input power and video output went through a camera cable connected to a wall socket, which was long enough to not cause any movement restriction while the participant was sitting down. The camera was connected to a multichannel digital video capture card in a recording computer in the room adjacent to the experiment room.

The head-mounted camera had a visual field of approximately 70°, horizontally and vertically. In a recent study, Yoshida and Smith [18] demonstrated the validity of this method for capturing the child's view of events. Using a similar context of tabletop play, they compared the direction of eye gaze as judged from frame-by-frame coding of the contents of the head camera to frame-by-frame coding of direction of eye gaze (recorded from a second camera fixed on the child's eyes). They found that 90% of head-camera video frames corresponded with these independently coded eye positions; the non-corresponding moments were brief, as well as rare (less than half a second). Their results indicate that at least, in the table top and toy play context, the contents of the head camera provide a good approximation of the visual information available to the child as a function of head and body movements.

C. Bird's Eye View Camera

A high-resolution camera was mounted right above the table and the table edges aligned with edges of the bird's eye image. As shown in Fig. 1 (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 is its high-quality video, which made the image segmentation and object tracking software work more robustly compared with the head-mounted mini camera that was light-weight, but with a limited resolution and video quality.

III. Experiment

A. Participants

We invited parents of toddlers in the Bloomington, IN, area to participate in the experiment. Fifteen children contributed data (six additional children were recruited, but either did not tolerate the head camera or were excluded because of fussiness before the experiment started). For the child participants included, the mean age was 21.3, ranging from 19.1 to 23.4 months. Ten of the included children were female and five were male. All participants were white and middle-class.

B. Stimuli

Parents were given a maximum of six sets of toys (three toys for each set) in a free-play task. The toys were either rigid plastic objects or plush objects with a simple and a single main color —factors that aided computalized automatic visual processing.

C. Procedure

Three experimenters conducted the study: one to distract the child; another to place the head-mounted camera on the child; and a third one to control the quality of the video recording. Parents were told that the goal of the study was simply to observe toy play and that they should try to interact with their child 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.

To calibrate the horizontal camera position in the forehead and the angle of the camera relative to the head, the experimenter placed a very attractive toy in front of the young child and signaled to a third experimenter in the adjacent room that the participant was clearly looking at the object and that the object was well centered. This outside experimenter, who controlled the recording, confirmed if the object was at the center of the image. If not, small adjustments were made on the head-mounted camera gear. After this calibration phase, the experimenters removed all objects from the table, asked the parent to start the experiment, and left the room. The parent was asked to take all three objects from one set, place them on the table, encourage the child to play with them, play with the child, and after hearing a command from the experimenters, remove the objects and bring out the next set of three. Each trial was 1 min long.

Not all of the 15 toddlers stayed in the experiment for the entire six trials (for example, some decided to leave the experiment after three or four trials by removing the head camera). Individual participants participated on average 3.7 trials (ranges 2–6). Across all 15 toddlers, a total of 56 trials were completed and these constitute the data analyzed in this study. The entire study, including initial setup, lasted for 10 to 15 min.

IV. Image Segmentation and Object Detection

The recording rate for each camera is 10 frames/s. Approximately 7200 ($10 \times 60 \times 6 \times 2$) image frames were collected from each dyad. The resolution of image frame is 720×480.

Visual information concerning the locations and sizes of objects, hands, and faces was automatically extracted from the raw camera images. Using computer vision techniques, this was accomplished in three major steps as illustrated in Fig. 2. Given raw images, the first step separates background pixels and object pixels. This step is not trivial in general because a first-person view camera continually moves causing moment-to-moment changes in visual background. In the present case, this step is helped considerably by the all-white background enabling the procedure to 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 is fixed in step 3).

The second step groups nonwhite pixels into several blobs using a fast and simple segmentation algorithm [19]. 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, as for example, when held in a participant's hand.

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 [20]. 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 as belonging to this specific object. Next, by putting probabilistic maps of all the possible objects together, and by considering the spatial coherence of an object, the detection algorithm assigns an object label for each blob in the segmented image.

The data derived from these steps for each frame are 1) the objects that are in the head-camera field; 2) the sizes of those objects in the field; and 3) whether a hand is holding an object (determined from the top-down view). An object is labeled as held by a participant if the object blob overlaps with a hand blob for more than 10 frames. We use this 1 s overlap requirement because we wanted to count *active* manual engagement with an object and not merely momentary overlap (in the camera image), as when a hand was passing by on the way to another object.

The validity of the automatic coding results were assessed by asking two human coders to annotate a small proportion of the data (~ 1200 frames); the comparison of these hand codings with the image processing results yielded 91% frame-by-frame agreement. One common disagreement in these frame-by-frame codings concerned holding that occurred when the hand was just above an object from the bird's eye view (but not holding that object), a fact easily seen by the hand coders; using the 1-s overlap rule in the automatic coding effectively eliminated this problem. Finally, we note the following results derive from cumulative statistics from thousands of image frames; therefore small errors in the automatic coding (image processing), errors are unlikely to change the overall patterns reported in the present study. In the next sections, we first report analyses of the contents of the head-camera images; we then report analyses relevant to hand movements and their role in selecting visual objects.

V. Visual Information Selection

Objectively, for all trials, there are three objects on the table and thus, three objects that could be in the child's view. These three objects are all approximately the same actual size and thus, when measured from the overhead camera, take up the same amount of area in the images from that camera. Moreover, if the child were to sit back and take a broad view of the table, not moving his or her head, all three objects would be in view in the head camera image and all would be approximately the same size. However, if the child moves his body and/or moves the objects so that one is closer to the head and eyes, than that selected object will be larger than the other objects, and being closer to the sensors could even obstruct the view of the other objects. If the child's head movements or manual actions on the objects focus successively on one more then another object, then the head camera images should show dynamic variation in the objects in view and in the relative sizes of those objects in the head camera view. In this way, the images in the head camera change as function of the child's own bodily movements and thus, provide data on how available information changes in real-time engagement with objects.

Fig. 3 shows frame-by-frame changes in the proportion of the head camera image taken up by each of the three objects (and also by the sum total of all body parts, faces, and hands from both participants, in yellow) from one trial of one child. This pattern is characteristic of all children. As is apparent, the number and size of the objects in the head camera image change frequently over the course of this trial. Sometimes only one object is in view, sometimes two or three. Moreover, the relative closeness of the objects to the head camera (and eyes), and thus the size of the objects in the head camera image, changes such that one object often dominates the image. The first set of analyses document this finding across the full data set which consists of 56 trials (from the 15 participating toddlers, 3.7 trials on average per toddler). All of the following measures and analyses are trial-based by averaging sensory data within a 60-s trial.

A. Number of Objects in View

Fig. 4 shows the average number of objects in the head camera image across all children and trials. All three objects are in the child's view less than 20% of time. Most often, there are only one or two (instead of three) objects in the head camera image (see also [18]). This is our first evidence that the child's visual field is selective.

B. Size of Objects in View

The size of the objects in the head camera view is a direct reflection of the closeness of the objects to the head (and head camera). The closest object—the one taking up most of the head camera field—seems likely to be the object being attended to by the child at the moment. To examine the degree and dynamic variation in which a single object dominates the head camera

image, we used several criteria (varying in their conservativeness) to define a dominating object in each image. Dominance was measured in two ways: 1) *absolute* size of an object, more specifically, the percentage of the head camera field that was taken up by the largest object in that field; 2) *relative* size of an object, the ratio of the dominating object to other objects in view. For the absolute size measure, as shown in the examples in Fig. 5, three increasing thresholds were used to define a frame as containing, or not containing, a dominating object: 3%; 5%; and 10% of the image (Given a 70° image size, the 3% criteria is roughly comparable to the size of the fovea). For the second relative measure, we used a ratio (largest object to the other two objects) of 0.50 for characterizing an object as "dominating" or not; a 0.50 ratio means that the larger object is at least larger than the *combination* of the other two objects.

As shown in Fig. 5, in more than 60% of frames, by the relative measure, there is one dominating object that is larger than the combination of the two objects (ratio > 0.5) and its absolute image size is also relatively distinct (3%).

By stricter absolute thresholds, almost 40% of time, there is a large object taking up at least 5% of the frame, and 10% of the time, there is a visually very large object taking up a substantial 10% proportion of the field. In brief, one of the three objects is often dominant in the field, which comprises another form of visual selection.

Because the head camera view is tightly tied—moment-to-moment—to the child's own actions, the dominating object may also change—in terms of its size, location, and relation to the other objects. Accordingly, we calculated the number of times that the dominating object changed, using the middle, 5% absolute threshold for defining a dominating object. By this measure, there are on average 12.8 switches in the dominating object per minute. This suggests frequent head and object movements, and thus rapid shifts in the information available. These rapid shifts, because they potentially relate to the child's own actions, may also be indicative of the child's momentary goals and the rapid switching of embodied attention in the cluttered environments of real world activity.

C. Discussion

Ordinary contexts in the real world are highly cluttered, with many objects and many potential targets for attention and for learning. Theorists of natural and artificial intelligence have often noted the daunting demands of attention in the "wild" [21], [22]. The results on the number and sizes of objects in child's visual field—as measured by the head camera—show that the child's view at any moment is often selective, limited to one or two of the three objects on the play table. Moreover, one of these objects is often closer than the others, and thus bigger in the field, dominating the view. Although the present analyses just demonstrate this fact, it could be crucial to building real-time models of how toddlers learn about objects and how they organize their attention. As shown in Fig. 3, the visual information from the child's head camera is highly dynamic: the three objects come in and out of view. The structure in this dynamic egocentric visual experience would seem essential to understanding embodied learning and, we note, it differs markedly in its very nature from the information captured from third-person cameras, the view on which most developmental research is based on.

The dynamic change in the objects in the head camera view and in their relative dominance in that view may be caused by several different actions: 1) the child's head rotation may change where the head camera is pointed; 2) the child may use their hands to bring objects closer to their head, and by so doing, may make the object in their hands much bigger and also make other objects smaller due to occlusion; and 3) the parent may also move objects close to the child's sensors (perhaps to attract the child's attention to that object). The results observed thus far, implicate the second two kinds of actions—hand actions—as the most likely major source of visual selection in the present study. This is because head rotation in general, though not

always, may dramatically change the locations of visual objects in view, but the distances between the head camera and all the object would not change. Therefore, head rotation will at most slightly increase or decrease the size of all the objects on the table. But hand movements literally can select one object to bring close to the head and eyes. Accordingly, the next set of analyses focus on hand movements.

VI. Hand Actions

A. Objects in Hands

In 68% of the frames, the child's hands are holding at least one of the three objects; in 55% of the frames, the parent is holding at least one object. Overall, in 86% of the frames, at least one participant is holding at least one object. These facts document the active manual engagement of the participants with the objects.

The main finding is that objects in hands determine the dominating object in the head camera image. Specifically, objects held by children are significantly larger (and thus more likely to be dominating by any of the criteria defined above). On average, the objects held by children take 4.5% of the child's visual field compared with the average size of objects in the image (3.2% of image). Objects in parents' hands are just slightly larger, 3.6%, than the average object size.

To affirm the statistically reliability of these differences, we compared object image sizes when the object was held by the child or by the parent to a control estimate of object image size—the average size all individual objects across all head-camera frames. This control measure provides the best estimate of average object size in this task context independent of hand action and actually *decreases* the likelihood of finding reliable differences between image sizes of hand-held objects as the control estimate includes the sizes of both the held and not held objects. An omnibus comparison shows that the image sizes of child-held, parent-held, and the control estimate of average object image size differ significantly, $F(2,165) = 109.36$, $p < 0.001$). Post-hoc comparisons (Tukeys hsd, all ps < 0.001) show that the image sizes of the objects in the child's hand were reliably larger than those in the parent's hands or by the control estimate. Objects in the parents' hands were also larger in the child's view than the control. In sum, both the child's and parent's actions effectively select objects for perception by influencing the objects that dominate in the child's visual field, however, the child's own manual actions play the more significant role in visual selection, at least in this toy-play task.

B. Moments With Dominant Objects

The next analyses provide converging evidence for these conclusions. Whereas the prior analysis started with objects in hands and then asked about their size in the head-camera image, the present analyses begin with head camera images and definitions of the dominating object in terms of the head camera image, and then ask whether those so-defined dominating objects are in the child's or parent's hands. As described earlier, the frame-by-frame definition of a dominating object may be based on both the absolute size of that object in the head camera view or its relative size with respect to other objects in view (the ratio of the largest to the sum of the other two). For these analyses, we report the two (middle and the highest) absolute size thresholds, 5% of the visual field and 10% of the visual field. The conclusions are qualitatively the same when based on the other measures as well. However, in these analyses, unlike the previous ones, we added a temporal stability criterion to the designation of the object as dominating, and the pattern of results we report below depend on this added criterion, which as we discuss later, may be important in its own right. The added requirement is that an object must maintain its dominance for at least 500 ms. That is, it must be at least somewhat *stably* dominant.

Based on these criteria, there are 10.2 dominant-object events per minute for objects with 5% dominance and 5.8 dominant-object events for objects with 10% dominance. In total, there are 2230 events with 5% dominance and 766 events with 10% dominance across all the subjects. The following results are derived from overall statistical patterns in these events calculated at the trial level.

At the 10% threshold (top of Fig. 6), dominating objects are in the child's hands more than 70% of time and much less often in the parent's hands (less than 20%), ($t(110) = 27.18$, $p < 0.001$). Indeed, as shown in the figure, parents were more likely to be holding *other* objects, not the one dominating at that moment in the head-camera image. At the 5% threshold (bottom plot in Fig. 6), the pattern is the same: the defined dominating object is more often in the child's than the parent's hands, $t(110) = 19.11$, $p < 0.001$.

These results suggest that the child's hand actions—and not the parent—play the key role in *visual* selection and indeed, from these analyses there is little evidence that parent hand actions play much of even a supporting role in this selection. As we discuss later, although parents clearly play an important role in such toy-play everyday interaction by introducing objects to the child, the results here suggest that visual selection is ultimately implemented by the child's manual actions. Given the rapidly changing views that characterize the head camera images (as evident in Fig. 3), stability may be the most critical attentional problem for toddlers. One conjecture is that the young perceiver's manual actions on an object actually stabilize attention, creating a perceptual-motor feedback loop centered on one object. We will revisit this topic in general discussion.

C. Moments Before/After an Object Dominates the Child's View?

Active toy play by toddlers, as Fig. 3 makes clear, generates rapidly changing head-camera views in which one and then another object dominates in the sense of being closer to the eyes, and thus bigger. What events in this dynamic lead up to some particular object *becoming* dominant in the image? There are at least four different kinds of behavioral patterns that could lead to a one-dominating object in the head-camera view: 1) the child's hands could select and move an object closer to their eyes; 2) the parent's hands could put the objects closer to the child; 3) the parent could move an object to the child, the child could then take it and move it close to the eyes, and 4) the child could move his or her body toward the table and probably also rotate the head toward one object to make that object dominate the visual field.

In an effort to better understand the dynamic processes that lead up to the dominance of some object in the head camera image, we zoomed into the moments just *before* and just *after* a dominating object became dominating and measured both the child's and the parent's behaviors. The approach is based on one used in psycholinguistic studies to capture temporal profiles across a related class of events [30] (in our case, the relevant class of events is a visually dominating object in the head camera image). Such profiles enable one to discern potentially important temporal moments within a trajectory. Fig. 7(a) and 7(c) shows the average proportions of time (which can be viewed as a probability profile) that a 10% or 5% "dominating object" was held by the child or parent. Thus Fig. 7(a) shows the probability that objects were in the child's or the parent's hands for the 10 s (10000 milliseconds) prior to a 10% threshold dominating object. The trajectory of the probability that the child was holding the to-be-visually-dominant object shows a clear and dramatic increase as function of temporal proximity to the visual dominance of the object. There is no such pattern for the trajectory of the parent holding the object. Indeed, if one assumes that the four possible hand states (holding one of three objects on the table or not holding) have an equal chance (that is 25%,) the probability that the parent is holding the to-be-dominant objects is close to chance and remains there as a function of proximity to the moment at which the object becomes visually dominant in the child's view.

One approach used in child development and psycholinguistic research asks when such trajectories first begin to diverge, which is commonly defined as the first significant difference in a series of ordered pairwise comparisons (see [23]–[26]). Ordered pair-wise *t*-tests of the child's and the parent's data reveal that these curves first diverge at around 7000 ms prior to dominance, $t(110) = 3.7$; $p < 0.001$. This thus defines the likely temporal window—and a long one—within which to explore in future work how perception-action loops may select and stabilize objects for vision.

Fig. 7(c) showed the same measure for the definition of the dominating object in terms of a 5% size in the head camera image. The pattern is similar, again showing an increasing probability that the to-be-dominant object is in the child's hand as a function of temporal proximity to the moment in which the object reaches the 5% threshold for visual dominance. However, by this more liberal criterion for visual dominance, the child and parent curves do not reliably diverge until 4000 ms prior to dominance ($t(110) = 6.76$; $p < 0.001$). This indicates that even with noisier data (resulting from the more liberal criterion for visual dominance), by 4 seconds prior to an object becoming visually dominant, the child's manual actions are already indicating its selection. Again, this analysis suggests the critical temporal window for future work directed to understanding how objects are selected *by the hand,* the role of visual events (in the periphery or generated by head movements perhaps) in causing objects to be *manually* selected, and then the unfolding events that lead to those objects being moved close to the head and eyes. At the very least, the present analyses make clear that the child's own actions play a strong role in visual selection in the sense of an object that dominates the child's view.

Fig. 7 also provides information on the role of the child's hands in terminating a dominant moment by making used-to-be-dominant objects less large in the head camera image. Figs. 7 (b) and 7(b) shows the results of this measure for the 10% and 5% thresholds. The parent's holding trajectories are between 20–25% which is again close to chance.

These conclusions were supported by separate child and parent analyses of the proportion of preceding and following trials that the dominant object was held. We conducted a 2 (threshold) by 2 (before or after) by 10 (1 s intervals) ANOVA of the data from individual children (there were 15 young participants in total). This analysis reveals reliable main effects of threshold, $F(1,560) = 56.26$, $p < 0.001$, of before and after $F(1,560) = 39.46$, $p < 0.001$, of time $F(9,560) = 123.67$, $p < 0.001$ and a reliable interaction between time and before and after, $F(9,560) = 25.63$, $p < 0.001$. The interaction is due to the symmetrical nature of the probability that the child is holding the to-be (before) and formerly (after) dominant object. This probability of holding increases prior to visual dominance, but then decreases post visual dominance. This analysis thus provides strong converging evidence for a strong link between the visual selection and children's manual actions on objects. The analysis of the parent data yielded no main effects or interactions that even approached conventional standards of statistical significance.

D. Discussion

The central contribution of this second set of analyses is that they tie visual selection (in the sense of objects close to the head and eyes) to the child's own manual actions. These results, of course, do not mean that *only* hand actions are important, (as compared to head and whole-body movements or to shifts in eye-gaze) but they do show that hand actions play a critical role in toddler visual attention, a role that has not been well studied. For the goal of building artificial intelligence and robotic systems, they also suggest the importance of building active sensors (cameras) that pan, tilt, and zoom, as well as, effectors that act on and move objects in the world, in both ways, changing the relation between the sensors and the effectors (see also [22]). Indeed, scholars of human intelligence often point to hands that can pick up and move

objects as central to human intelligence, linking manual dexterity to language [27], to tool use [28], and to means-end causal reasoning [29]. The present results hint hands and their actions on objects may also play a role in organizing visual attention, at least early in development. We do not know from the present study how developmentally specific the present pattern is, whether it generally characterizes all of human active vision or whether it is most critical and most evident within a certain developmental period. This is an important question for future research.

From the present data, we also do not know what instigates the child's manual grasp of an object. These could start with a rapid shift in eye-gaze direction (or head movement) that then gives rise to reaching for the object and bringing it close [30]. That is, manual action may not be the *first* step in selection, but rather may be critical to stabilizing attention on an object. In this context, it is worth noting that although the present findings indicate that it is the child's hand actions that play the more critical role in making some objects dominant in the head-camera image compared to the parent's hand actions, this does not mean the parent's actions play no role. Indeed, a parent's touch to or movement of a non-dominating object in the child's view could start the cascade—of look, grasp, and sustained attention. Future comparisons of the dynamics of attention of children engaged with toys and with a mature social partner versus when playing by themselves may provide critical data on this point.

A final open question in these analyses is the possibility of individual differences. Casual observation (as well as a robust developmental literature, [31]–[34]) suggests that some parents play a more active role in directing attention to individual objects than others. The analyses in this section were all based on group data and so provide no information on this issue; however, examination of the probability of holding prior to a visual dominating event by parent-child dyad did not reveal clear individual differences. This could be due to the lack of sufficient data from any individual dyad to discern these patterns or from the possibility that the key individual differences reside in what starts the process, but not in the attention sustaining processes of reaching to and holding an object. That is, once the child's attention is directed (one way or the other, either through the parent's successful bid for the child's attention or through the child's self-generated selection), the role of body and hands may remain the same.

VII. General Discussions, Limitations, and Conclusions

The fact that young children's own bodily actions create their visual experiences is well recognized [9], [35], [36]. In addition, there is now a growing literature on infant and children's eye movements and their role in learning about objects [37]–[39]. However, there has been little study of how larger bodily movements—of head, posture, hands—structure and select visual experience. The present results strongly suggest that there are insights to be gained by taking a larger whole-body approach to visual attention, at least if the goal is understanding attention in actively engaged toddlers. To this end, the use of a head camera that moves with the child's movements and that captures the objects directly in front of the child's face provides a way of capturing the dynamic coupling of vision and body movements beyond shifts in eye gaze.

Overall, the results strongly implicate manual activity (at least for toddlers in the context of toy play) in selecting, and perhaps also stabilizing visual information. As children use their hands to bring objects of interest close to the face, those objects increase in their visual size and also block the view of other objects. These actions and their consequences for vision, mundane as they might seem, naturally segment and select objects in a cluttered visual field. Thus, they may prove to be important ingredients in toddler intelligence and learning. Indeed, the natural consequences of bodily action on the available visual information may be crucial to achieving human-like prowess given noisy and ambiguous data, providing a peripheral (and

perhaps cognitively "cheaper") solution to visual selection. Developmental theorists in the past have often sought to solve the noisy input problem through innate constraints (e.g., [13]) or, more recently, through the actions of the social partner as an orchestrator of the child's attention (e.g., [15]). The present results do not necessarily diminish the role of conceptual constraints in some learning tasks nor the role of the mature social partner, but they do show that in everyday tasks of acting and playing with objects, children's own hand actions may be a key part of the process.

The role of hands in bringing objects close to the eyes and thus in increasing their dominance in the visual field raises a number of interesting developmental questions. Infants do not intensively manually play with and explore objects for sustained periods until they sit steadily [37], which occurs around 8 months. This suggests possibly important developmental changes in visual attention, object segmentation, and selection after this period, a conjecture for which Soska *et al.* [35] have already presented some preliminary evidence. Their findings, in conjunction with the present results, point to the importance of future work that takes a more developmental approach by examining how attention, sustained attention to visual objects, and manual action on objects, changes in the first two years of life–a period of dramatic change in what infants manually do with objects (see also [40]).

Before concluding, the limitations of the present method also warrant discussion. One contribution of the present approach is the use of the head camera which provides profoundly different information about the toddler's activities in the task than does a third person camera, which is the standard approach used in child development research. The difference between a head camera and a third person camera is that the first person camera captures the momentary dynamics of available visual information *as it depends* on the child's own actions. The limitation, however, is that not all actions influence the head camera view; in particular, the head camera moves with head movements, not eye movements. Thus, the head camera is *not* a substitute for direct measures of eye gaze direction [18], [41], but instead provides information about the dynamics of *available* visual information with larger body movements. In the ideal, one would jointly measure both the dynamics of the larger visual field (as given by a head camera), and also focal attention as indicated by eye-gaze direction within that field. Recent advances in developmental research (e.g., [42]) may make this possible in the near future.

A second limitation of the work concerns the definition of the dominant object in the head camera image. An object that is very large in the visual field—that the child has brought close to their own face—has considerable face-validity for being the object being attended to. However, given that there has been no prior work in this area, it is unclear just how big an object needs to be in a head camera field to count as dominating attention. It was for this reason that we used multiple and converging measures. A next step needed to validate this approach is to link the dominating object, as measured here, to some other behavioral outcome related to attention, for example, to learning about the object or to ease the distraction by some other salient objects in the periphery.

To conclude, the results reported here have the potential to contribute to understanding human intelligence and to building autonomous intelligence in several important ways. First, they emphasize the role of the child's own actions. Considerable recent research on both human and artificial systems has focused on the social context and how the parent selects information by guiding the child's attention (e.g., [19], [29]). Indeed, recent robotic studies clearly demonstrate how artificial devices with impoverished motor and effector systems may be bootstrapped to intelligent behaviors through the actions of a human partner [43]–[45]. But these demonstrations may be missing the other key part of autonomous intelligence—self-generated actions on the world and the self-organizing properties of perception-action loops.

The present results make clear that, at least for the age group in this study, the child's own activity is a key component in organizing visual input.

Second, the results strongly 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 seriously benefit aspects of object recognition (including segregating objects, integrating views, and binding properties, see [9]). The central importance of hand activities as they relate to language [46], social communication[47], [48], tool use [28], and problem solving [29] are well noted and central to many theories of the evolution of human intelligence [8]. They may also be foundationally relevant to human-like visual attention to objects.

Third, our approach introduces a new method, with many advantages (despite some already noted limitations). Most studies of child activity use a third-person view camera to record the whole scene as viewed by an outside observer. Thereafter, researchers observe and analyze recorded video clips to identify interesting behavioral patterns. A human coder's observations are necessarily influenced by the larger structure of the scene and their adult interpretation of its structure. The approach used here differs in two important ways from the usual techniques in child development research 1) by using a head-mounted camera, we capture the dynamic first-person view; 2) by using automatic image processing and coding methods, the data are both more objective and more fine-grained than usual behavioral coding schemes.

Fourth, all of these contributions are relevant to building smarter 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. Indeed, considerable advances have been made in biologically inspired forms of artificial intelligence (e.g., [4], and [49]–[52]) and there is a growing realization that a deeper understanding of how human children *develop* may provide the best clues for building human-like intelligence [7], [8], [53], [54].

If we were to offer engineering suggestions from what we have learned from this study of toddler's visual attention during toy play, they would be this embodied solution: build a device with hands that can reach out, hold, and move objects in the world, and that brings those objects, one at a time, close to the sensors.

Acknowledgments

The authors wish to thank Amanda Favata, Amara Stuehling, Mellissa Elston, Andrew Filipowicz, Farzana Bade, Jillian Stansell, Saheun Kim, and Mimi Dubner for collection of the data. They would also like to thank the Associate Editor Gedeon Deak and three anonymous reviewers for insightful comments.

This work was supported in part by National Science Foundation Grant BCS0544995 and by NIH grant R21 EY017843. The work of A. F. Pereira was also supported by the Portuguese Fulbright Commission and the Calouste Gulbenkian Foundation.

References

1. Weng J, et al. Artificial Intelligence: Autonomous Mental Development by Robots and Animals 2001;5504:599–600.
2. Yu C, Ballard D, Aslin R. The role of embodied intention in early lexical acquisition. Cogn Sci: Multidisciplinary J 2005;29(6):961–1005.
3. Deák G, Bartlett M, Jebara T. New trends in cognitive science: Integrative approaches to learning and development. Neurocomputing 2007;70(13–15):2139–2147.

4. Gold, K.; Scassellati, B. A Robot That Uses Existing Vocabulary to Infer Non-Visual Word Meanings From Observation. p. 883
5. Oates, T.; Eyler-Walker, Z.; Cohen, P. Toward Natural Language Interfaces for Robotic Agents: Grounding Linguistic Meaning in Sensors. p. 227-228.
6. Roy D, Pentland A. Learning words from sights and sounds: A computational model. Cogn Sci: Multidisciplinary J 2002;26(1):113–146.
7. Asada M, et al. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. Robot Auton Syst 2001;37(2–3):185–193.
8. Pfeifer, R.; Scheier, C. Understanding Intelligence. Cambridge, MA: MIT Press Cambridge; 1999.
9. Metta G, Fitzpatrick P. Better vision through manipulation. Adaptive Behavior 2003;11(2):109.
10. Bertenthal B. Origins and early development of perception, action, and representation. Annu Rev Psychol 1996;47(1):431–459. [PubMed: 8624139]
11. Asada M, et al. Cognitive developmental robotics: A survey. IEEE Trans Auton Mental Develop 2009;1(1):12–34.
12. Rolf M, Hanheide M, Rohlfing K. Attention via synchrony: Making use of multimodal cues in social learning. IEEE Trans Auton Mental Develop 2009;1(1):55–67.
13. Spelke E, Kinzler K. Core knowledge. Develop Sci 2007;10(1):89–96.
14. Quine, W. Word and Object. Cambridge, MA: MIT press; 1964.
15. Baldwin, D. Joint Attention: Its Origins and Role in Develop. 1995. Understanding the link between joint attention and language; p. 131-158.
16. Land M, Hayhoe M. In what ways do eye movements contribute to everyday activities? Vision Res 2001;41(25–26):3559–3565. [PubMed: 11718795]
17. Hayhoe M, Ballard D. Eye movements in natural behavior. Trends Cogn Sci 2005;9(4):188–194. [PubMed: 15808501]
18. Yoshida H, Smith L. Hands in view: Using a head camera to study active vision in toddlers. Infancy. 2007
19. Comaniciu, D.; Meer, P. Robust Analysis of Feature Spaces: Color Image Segmentation. p. 750-755.
20. Pentland, A.; Moghaddam, B.; Starner, T. View-Based and Modular Eigenspaces for Face Recognition. p. 84-91.
21. Breazeal C, Scassellati B. Infant-like social interactions between a robot and a human caregiver. Adapt Behav 2000;8(1):49.
22. Ballard D, et al. Deictic codes for the embodiment of cognition. Behav Brain Sci 1997;20(4):723–742. [PubMed: 10097009]
23. Gershkoff-Stowe L, Smith L. A curvilinear trend in naming errors as a function of early vocabulary growth. Cogn Psychol 1997;34(1):37–71. [PubMed: 9325009]
24. Thomas M, Karmiloff-Smith A. Are developmental disorders like cases of adult brain damage? Implications from connectionist modelling. Behav Brain Sci 2003;25(6):727–750. [PubMed: 14598624]
25. Tanenhaus M, et al. Integration of visual and linguistic information in spoken language comprehension. Science 1995;268(5217):1632–1634. [PubMed: 7777863]
26. Allopenna P, Magnuson J, Tanenhaus M. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. J Memory Language 1998;38(4):419–439.
27. Pollick A, de Waal F. Ape gestures and language evolution. Proc Nat Acad Sci 2007;104(19):8184. [PubMed: 17470779]
28. Lockman, J. Child Develop. 2000. A perception-action perspective on tool use development; p. 137-144.
29. Goldin-Meadow S. Beyond words: The importance of gesture to researchers and learners. Child Develop 2000:231–239. [PubMed: 10836578]
30. Jeannerod M, et al. Grasping objects: The cortical mechanisms of visuomotor transformation. Trends Neurosci 1995;18(7):314–320. [PubMed: 7571012]

31. Tomasello M, Farrar M. Joint attention and early language. Child Develop 1986:1454–1463. [PubMed: 3802971]
32. Tomasello M, Todd J. Joint attention and lexical acquisition style. First Language 1983;4(12):197.
33. Pereira A, Smith L, Yu C. Social coordination in toddler's word learning: Interacting systems of perception and action. Connection Sci 2008;20(2–3):73–89.
34. Nagai Y, Rohlfing K. Computational analysis of motionese toward scaffolding robot action learning. IEEE Trans Auton Mental Develop 2009;1(1):44–54.
35. Soska K, Adolph K, Johnson S. Syst. in Develop.: Motor Skill Acquisition Facilitates Three-Dimensional Object Completion. unpublished.
36. Ruff H. Components of attention during infants' manipulative exploration. Child Develop 1986:105–114. [PubMed: 3948587]
37. Johnson S, Amso D, Slemmer J. Development of object concepts in infancy: Evidence for early learning in an eye-tracking paradigm. Proc Nat Acad Sci 2003;100(18):10568–10573. [PubMed: 12939406]
38. Johnson S, Slemmer J, Amso D. Where infants look determines how they see: Eye movements and object perception performance in 3-month-olds. Infancy 2004;6(2):185–201.
39. von Hofsten C, et al. Predictive action in infancy: Tracking and reaching for moving objects. Cognition 1998;67(3):255–285. [PubMed: 9775511]
40. Smith L. From fragments to shape: Changes in human visual object recognition between 18 and 24 months. Current Direct Psychol. unpublished.
41. Aslin R. Headed in the right direction: A commentary on Yoshida and Smith. Infancy 2008;13(3): 275–278.
42. Adolph KE, et al. Head-mounted eye-tracking with children: Visual guidance of motor action. J Vision May;2008 8(6):102.
43. Breazeal, C. Designing Sociable Robots. Cambridge, MA: The MIT Press; 2004.
44. Brooks R, et al. The cog project: Building a humanoid robot. Lecture Notes Comput Sci 1999:52–87.
45. Scheutz, M.; Schermerhorn, P.; Kramer, J. The Utility of Affect Expression in Natural Language Interactions in Joint Human-Robot Tasks. p. 226-233.
46. Bates E, Dick F. Language, gesture, and the developing brain. Develop Psychobiol 2002;40(3):293–310.
47. Bakeman R, Adamson L. Infants' conventionalized acts: Gestures and words with mothers and peers. Infant Behavior & Develop 1986;9(2):215–230.
48. Carpenter M, et al. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. Monogr Soc Res in Child Develop. 1998
49. Steels, L.; Vogt, P. Grounding Adaptive Language Games in Robotic Agents. p. 474-482.
50. Steels L, Kaplan F. AIBO's first words: The social learning of language and meaning. Evol Comm 2001;4(1):3–32.
51. Yu C, Ballard D. A multimodal learning interface for grounding spoken language in sensory perceptions. ACM Trans Appl Percep 2004;1(1):57–80.
52. Billard A, et al. Discovering optimal imitation strategies. Robot Auton Syst 2004;47(2–3):69–77.
53. Smith L, Gasser M. The development of embodied cognition: Six lessons from babies. Artif Life 2005;11(1–2):13–29. [PubMed: 15811218]
54. Smith L, Breazeal C. The dynamic lift of developmental process. Develop Sci 2007;10(1):61–68.

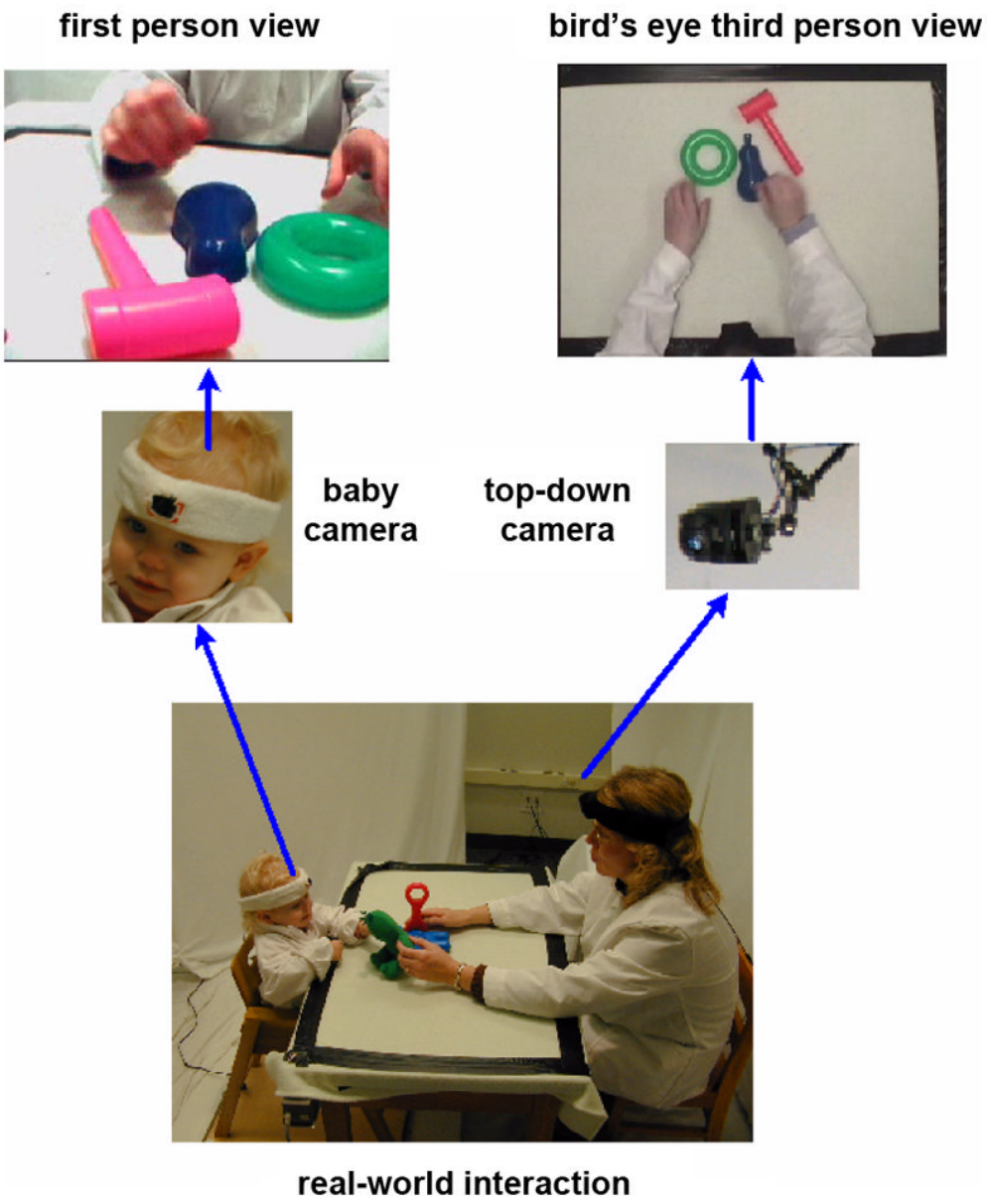

Fig. 1.
Multicamera sensing system. The child and the parent play with a set of toys at a table. A mini-camera is placed onto the child's head to collect visual information from a first-person view. Another camera mounted on the top of the table records the bird's-eye view of the whole interaction.

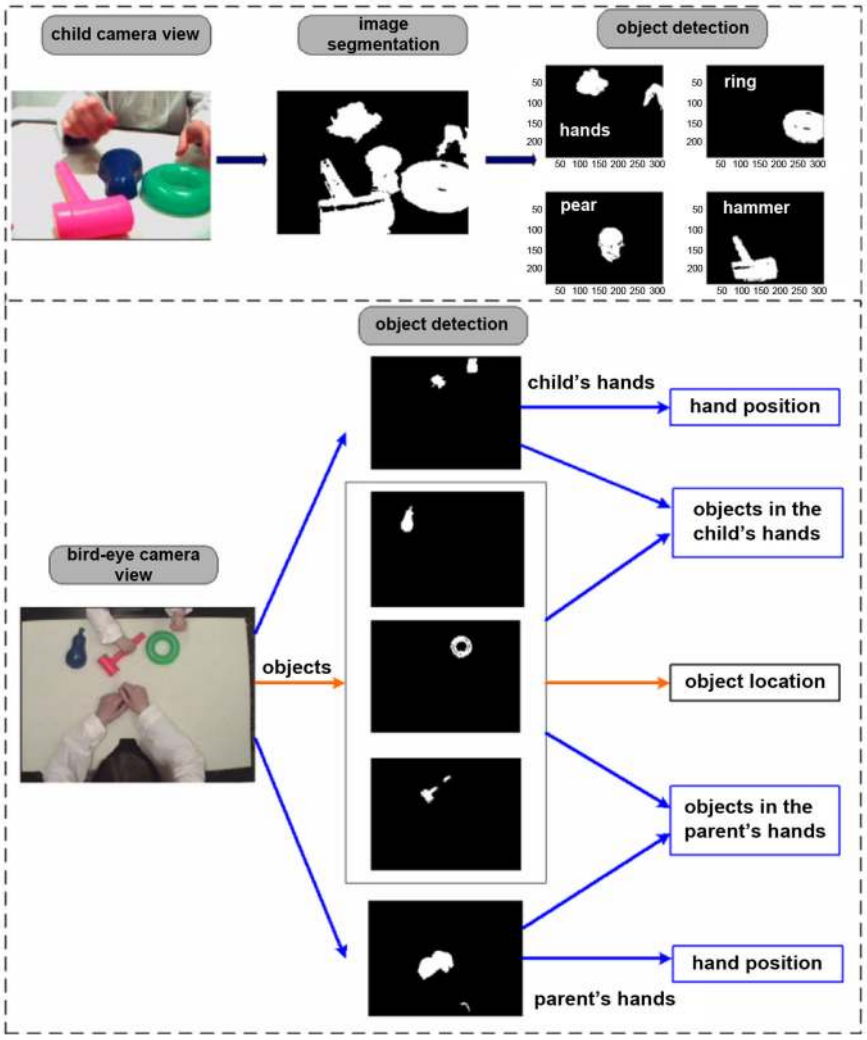

Fig. 2.

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's eye view camera. The information about whether a child or a parent holds an object is inferred based on spatial proximity of a hand blob and an object blob from a third-person view.

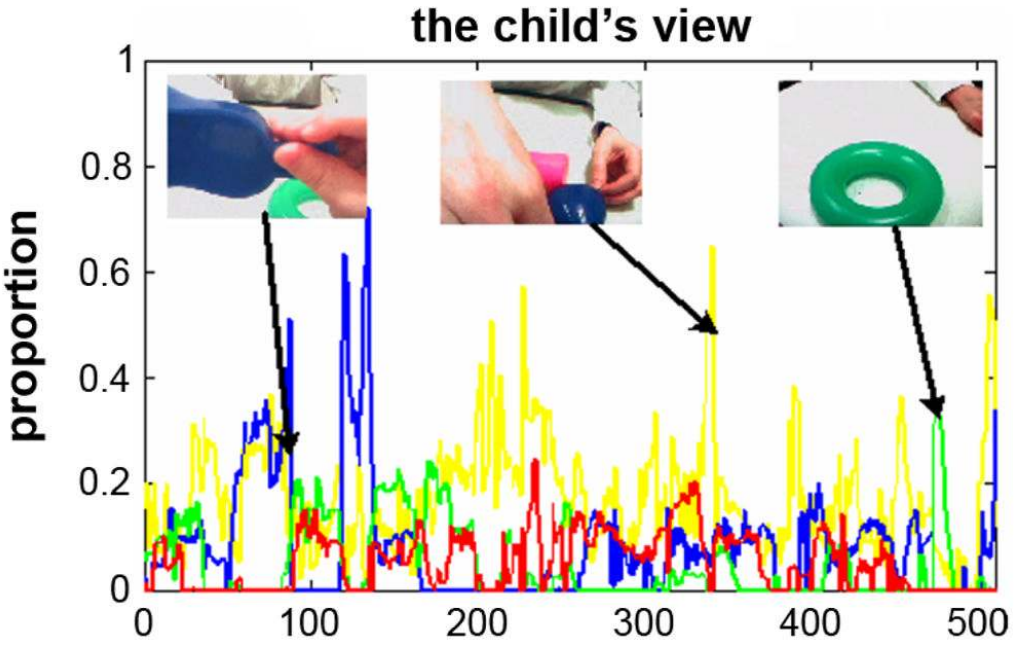

Fig. 3.

The proportions of visual objects in the first person view. X-axis represents image frame numbers. Each trajectory represents the proportion of one of three objects or hands in the child's visual field. The results show that the child's visual field is very dynamic and very narrowly focused on attended object at the moment.

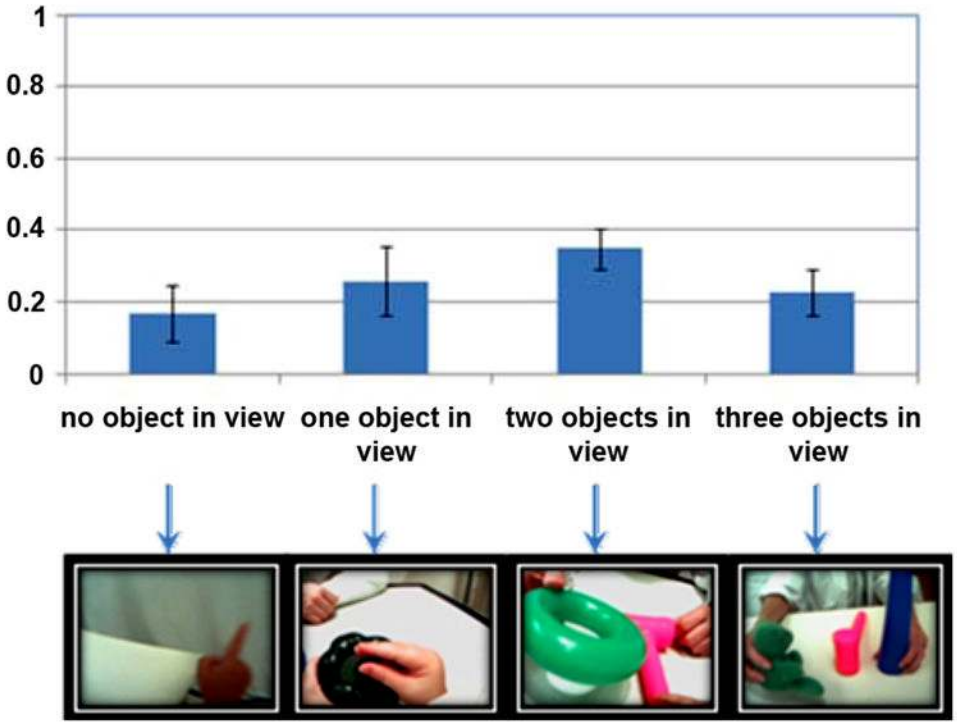

Fig. 4.

The proportion of time that objects are in the child's visual field. Note that although there are always three objects on the table, for 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. A dominating object is defined based on both the absolute size of an object and its relative size with other objects in view.

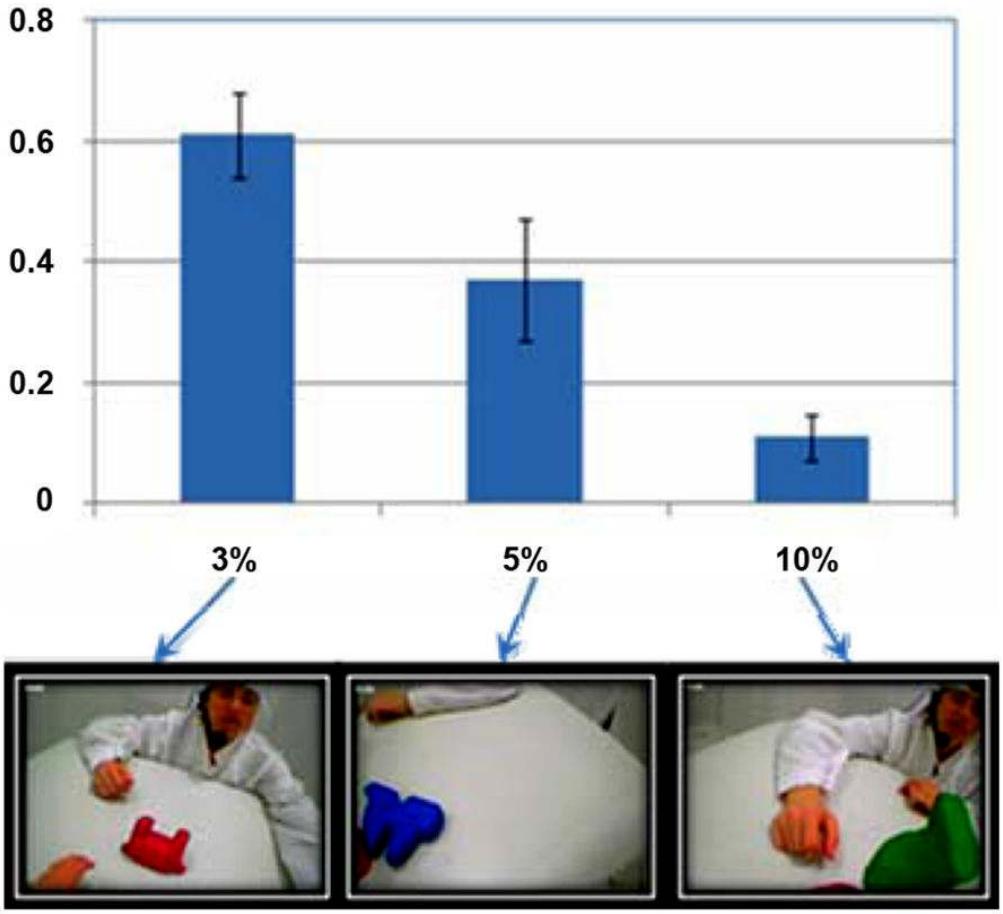

Fig. 5.

The proportion of time that there is always a dominating object in the child's visual field at a moment. A dominating object is defined based on both the absolute size of an object and its relative size with other objects in view. Three absolute object sizes (3%, 5%, and 10%, etc.) are used to be combined with a fixed relative ratio 0.5.

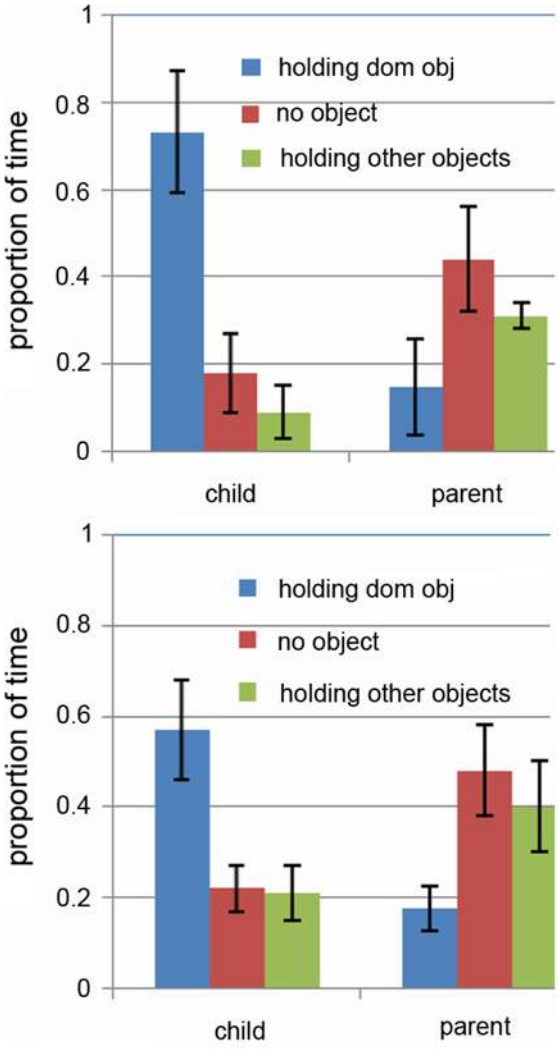

Fig. 6.

The proportion of time that the child's or the parent's hands hold a dominant object in those dominant moments. Top: Dominant moments with 10% dominance. Bottom: Dominant moments with 5% dominance.

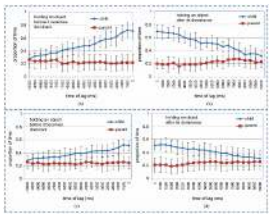

Fig. 7.

(a) and (c): The proportion of time that either the child or the parent is holding a to-be-dominant object. Plot (a) shows the results from 10% dominance events and Plot (c) are derived from 5% dominance events. (b) and (d): The proportion of time that either the child or the parent is holding an used-to-be-dominated object. Plot (b) is derived from 10% dominance events and Plot (c) is from 5% dominance events.