Understanding Embodied Visual Attention in Child-Parent Interaction

Sven Bambach
School of Informatics and Computing
Indiana University
Bloomington, IN
sbambach@indiana.edu

David J. Crandall
School of Informatics and Computing
Indiana University
Bloomington, IN
djcran@indiana.edu

Chen Yu
Psychological and Brain Sciences
Indiana University
Bloomington, IN
chenyu@indiana.edu

***Abstract*—A key component of the human visual system is our attentional control — the selection of which visual stimuli to pay attention to at any moment in time. Understanding visual attention in children could yield new insight into how the visual system develops during formative years and how their visual attention and selection play a role in development and learning. We use head-mounted cameras to record first-person video from interacting children and parents, giving a good approximation of the contents of their visual fields of view, and collect gaze direction data to record where they look within the visual field. We data-mine this data to study the distributions of gaze patterns within the first-person visual frame for both children and adults. We also study the ability of visual saliency to predict visual attention, as a function of the tasks, actions, and interactions that the participants perform. We find significant differences in the results between children and parents, indicating substantial differences in how their bodily actions are coupled with their visual attention between developing (child) and developed (adult) visual systems.**

I. Introduction

The visual world is cluttered with targets and distractions, requiring us to select and stabilize attention on just a subset of visual information in real time. For this reason, our visual system actively searches for relevant information from our environment on a moment-by-moment basis, by for example generating 3 to 5 saccades per second [1]. Our attentional control involves two inter-related modes: (1) a voluntary, goal-directed mode, in which attention is guided by contextually appropriate goals and intentions, and (2) an involuntary, stimulus-driven mode, in which attention is captured by physically salient stimuli. In addition, there is growing evidence that the brain is optimized to learn perceptual stimuli that signal the potential for procuring reward, exerting an additional influence on attentional deployment [2], [3].

Work to understand visual attention has typically been conducted in well-controlled experimental settings. With recent advances in sensing and computing techniques, however, it has become possible to use lightweight head-mounted cameras to record visual information from a first person perspective, and then to data mine this egocentric video data. The view from a forehead-mounted camera close to the eyes roughly approximates the visual field of a person. This new paradigm, providing a personal view of the world, opens up unique opportunities in understanding human vision systems [4], [5], [6], as well as innovative consumer products like Google Glass.

But with these opportunities for egocentric views come challenges, notably because first-person video is very dynamic compared to video from a stationary camera. Since a first-person camera is attached to the head, every head turn and every change in body orientation causes global changes in the first-person view. Some of these changes are task-relevant and goal-directed, while others are spontaneous. From a computer vision perspective, most existing algorithms for object recognition and tracking assume cameras are stationary or have a known, simple motion model. From a cognitive perspective, large and rapid head movements challenge attention by changing the visual information available to the sensors and by disrupting the alignment of body-centric spatial frames of reference for directed action and for orientation.

Our aim in this paper is to understand visual selection in freely moving toddlers, in tasks with changing goals, competing visual targets, and dynamic visual information that changes with the child's own actions. These are challenging conditions for both computer vision and human cognition, but are the actual contexts in which toddlers learn in the real world. We have three primary goals in this paper. First, we study where children and adults look in their first person views by analyzing eye gaze data synchronized with video from head-mounted cameras. Second, to better understand visual attention mechanisms, we use saliency maps to predict where people look, studying the contexts in which saliency accurately predicts gaze and those in which it does not. Third, we compare these analyses of gaze distribution and the predictive power of saliency between children and parents. To our knowledge, our study represents the first attempts to computationally analyze high-density eye gaze, head, and hand movement data with children's first-person video, and to document micro-level behavioral patterns that link bodily actions and visual saliency to gaze direction in first-person views.

II. Related work

Since Posner's classic paper on attention as a spatial spotlight [7], we have learned a great deal about the importance of localized attention for selection [8]. Experiments show that adults can readily attend to one specific location (and more rapidly detect objects at that location) without moving the eyes and while eye gaze is fixated elsewhere [9]. Thus spatial attention in adults can be internal, not requiring moving the sensors toward the attended object. For example, Dorr *et al* [10] study gaze patterns on various kinds of stimuli, including

static images and different types of videos like Hollywood and natural films. They find significant differences in gaze patterns between these stimuli, suggesting that human attention differs significantly depending on context and stimuli type.

However, attention is also tied to the body: adults typically orient eye gaze to the attended location, while eye movements [11], [12], head movements [13], and even hand movements [14], [15] bias visual attention in the direction of the movement. Visual attention thus appears to be coupled to mechanisms of directional action, perhaps because we often direct attention in preparation for action. Along these lines, Tatler *et al* [16] demonstrate the limits of visual salience in predicting eye gaze, finding that while bottom-up saliency measures are highly predictive of gaze in static images, they do not generalize well to naturalistic everyday contexts with arbitrary viewing behavior. They suggest that models need to consider other higher-level factors including learned prior information about the environment, uncertainty in visual observations, and the specific task at hand. Indeed, using head-mounted eye trackers became a recent trend in studying visual attention of adults in the context of natural behaviors, which has already led to important findings [17]. However, due to technical limitations (like camera weight), there have been few studies on tracking and recording first-person view video and eye movements until recently [5], [6], [18]. The present study represents our most recent efforts at the frontier of this new venue by analyzing gaze data and visual saliency at the pixel level from video data captured from first-person view cameras.

In parallel, computer vision researchers have begun to consider first-person video streams, driven in part by emerging consumer egocentric cameras like Memoto and Google Glass. The rapid and unpredictable camera motion of head-mounted cameras creates challenges compared to more traditional video from stationary cameras. Recent work on egocentric video has included video summarization [19], recognizing objects [20], and inferring the user's actions from object interaction [21] and eye gaze [22]. While our paper is not about computer vision, we believe that our studies of saliency, attention, and gaze prediction in first-person video streams provide both empirical evidence and useful insights on better models that may improve these computer vision algorithms in the future.

III. Methods

To realize our goal of understanding visual attention in first-person views, we developed a system that captures a variety of video and sensing data from interacting participants in a lab. We then use image processing algorithms to automatically extract fine-grained coded data from these raw sensor feeds.

A. Multimodal sensing system

Our sensing environment monitors parents and children engaged in free-playing interaction with several toy objects, as shown in Figure 1. Using this environment, we measured the moment-to-moment body positions of participants via head- and hand-mounted sensors with a Polhemus 6 Degree-of-Freedom (DOF) motion tracking system. In addition, we captured each participant's visual field via tiny cameras on head-mounted eye trackers. The angle of the camera is adjustable and has a visual field of about 90° horizontally. Each eye

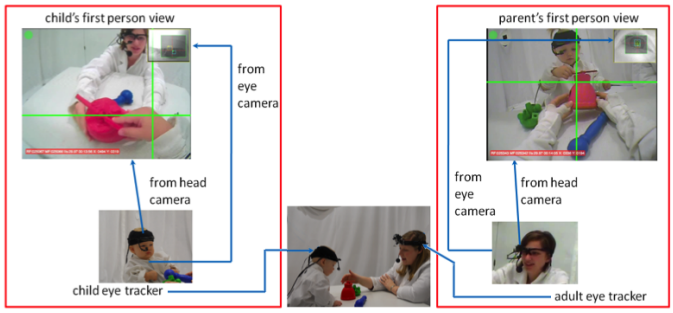


Fig. 1. Experimental setup. We use 4 cameras to record joint play between child and parent: 2 head-mounted cameras (part of the eye-tracking system worn by each participant) record data from the perspective of the participant, an overhead camera sees the table and the hands of both the infant and parent, and a scene camera records both people as if watched by an observer. High temporal resolution (30Hz for first-person view and gaze tracking, 120Hz for head and hand movements) allows detailed measurement of the temporal dependencies within and between the sensorimotor systems of the participants.

tracker also monitored where a person is looking (the green crosshairs in Figure 1), through an infrared camera pointing at each participant's left eye. (The validity of this method was demonstrated in [6]). In addition to the two head-mounted eye trackers (on the parent and the toddler), a third camera situated over the table recorded a bird's eye view of the interaction, and a fourth camera viewed the whole scene from the side.

Using this experimental setup we collected data from 13 child-parent dyads (with a success rate in placing sensors on children of over 70%). The mean age of toddlers is 13 months with a standard deviation of 3.2 months. Parents were simply told to engage the child with the toys and otherwise interact as naturally as possible, leading to a free-flowing interaction with no constraints on where parents or children should look, or what they should do or say. Toddlers and parents typically exchanged objects back and forth, engaged in joint actions with the objects, took turns and shifted attention. There were two interaction trials, each lasting about 2 minutes, in which the participants were given different sets of three toy objects. A typical study (including setup) lasted 10 to 15 minutes, from which we collected about 15 gigabytes (GB) of data from each dyad. Our analysis in this paper focuses on a subset of this data: egocentric video, eye tracking, and body movements.

As one of the very first studies using head-mounted eye tracking on toddlers, we note two limitations of the experimental setup. First, the human visual field is much broader (190° for adults) than the visual angle of the first-person view camera (90°). Nonetheless, as demonstrated by previous studies [17], well-calibrated first-person video and eye movement data are still reliable approximations of people's visual fields and what they attend to in the view. Second, the interaction environment in our lab setting is much less cluttered than the real world, since we cover the background with white curtains. We do this to occlude task-irrelevant distractors so that participants focus on free-play, attending solely to the toys and the other's face which are the regions-of-interest (ROIs) in the study.

B. Data Analysis

1) Video data processing: The recording rate for each of the four cameras (overhead, side observer, and two head-mounted cameras) is 30 frames per second with a spatial

resolution of 720×480 pixels. We conducted two forms of image processing on the image frames of these videos:

1. Pixel-level visual saliency estimation. We used the seminal saliency map model of Itti *et al* [23] to estimate which areas of the image were most salient, according to motion, intensity, orientation, and color cues. Itti's saliency model applies bottom-up attention mechanisms to topographically encode conspicuity (or "saliency") at every location in the visual input. These analyses give a description of the first-person view over time in terms of the visual properties that might be relevant to stabilizing or shifting attention.

2. Object segmentation and recognition. We also extracted visual information at a higher semantic level, in particular estimating the locations of objects, hands, and faces from the videos using the computer vision techniques detailed in prior work [5], yielding the position of the specific objects in view. This analysis also estimates relative object sizes (which vary as objects are brought closer to the eyes). We then combined this with momentary gaze data to detect which object was attended to in the first person view. In addition, we manually annotated object holding activities for both children and parents.

2) Motion data processing: Six motion tracking sensors were placed on participants' heads and hands to record 6 degree-of-freedom measurements (3 translational dimensions plus 3 rotational dimensions) of their head and hand movements at a frequency of 240 Hz. Our primary interest in this paper is the overall dynamics of body movements. We thus computed magnitude of positional changes (by computing magnitudes of first derivatives along the translational dimensions) and orientation change (along the rotational dimensions). We found that head position movements are equally frequent in children and parents, but that children rotate their heads much more frequently than adults do.

IV. Experimental Results

Having described our experimental setup and data analysis methods, we now turn to reporting our results. Our experiments had two main goals: (1) to characterize where people look within their first person view, (2) to study the connection between visual saliency and eye gaze (attention) in first-person views. We study these goals in both children and parents, and under a variety of contexts and conditions.

A. Eye gaze distributions

Eyes and head movements are known to be well coordinated because people tend to adjust their head orientation to point in the same direction as eye gaze [24]. However, few studies investigate gaze direction in egocentric views in the context of natural behaviors (but also see [6]). As one of the first experiments to record gaze data from head-mounted eye trackers in free-flowing interaction, we first report where children and parents look within the first person views.

Figure 2(a) visualizes these results using heatmaps of gaze distribution in the first-person view for both parents and children (generated using the technique of [25] with smoothing parameter $\lambda = 50$). Even with 13 dyads, the high-density measurements in our experiments provide tens to hundreds of thousands of gaze observations (indicated by N in Fig 2). As expected, gaze is most often in the center of the visual field for both children ($\mu = (340, 231)$ where the center of the visual field is $(360, 240)$ in our coordinate system) and parents ($\mu = (361, 224)$). Even though this is not surprising, the quantitative results reported here are still informative to applications. For example, if we know when gaze is likely to be centered in the first person view, we can better build computer vision systems to estimate visual attention from egocentric views, and better understand human eye-head coordination. More interesting, however, is that toddler gaze is more spread out around the center of the visual fields ($\sigma_x = 86, \sigma_y = 65$, where σ_x is the horizontal standard deviation of gaze points and σ_y is the vertical standard deviation), whereas parent gaze was much more compact ($\sigma_x = 53, \sigma_y = 60$). This suggests that adults' developed visual systems are better able to coordinate eye and head movements than toddlers, whose visual systems are still developing.

Two primary behavioral activities may jointly cause both the dynamics of the first-person view and the visual attention switches in the first person view. Head turns change what a person sees; when one turns his head without a gaze shift, he sees a new spatial location even though the eye-in-head position has not changed. In contrast, if one wants to fixate on the same spatial location in the world, then one has to adjust gaze direction to compensate for the head turn. Second, manual manipulation of objects may bring objects closer to or further away from the eyes. Even given a stable visual field (with stationary head position), these manual activities dramatically change what one sees as well as what one attends to. Given these observations, we next calculated heatmaps for a subset of video frames in several specific contexts: whether or not the head was moving, whether the person was looking at an object or a face, and whether the person was holding an object.

Head motions. Figure 2(b) splits the heatmaps from Figure 2(a) into the moments when the head was stationary versus when it was moving. The results show that for both children and parents, eyes and head are more well-coordinated when the head is stable. Moreover, gaze-in-head direction in parents is more clustered around a central location and that location is slightly above the center of the visual field (at $(374, 210)$ when stationary, compared to the true center point of $(360, 240)$). For children, gaze-in-head direction when the head is stable is closer to the center of the visual field $(335, 226)$ and more spread out ($\sigma_x = 84$, $\sigma_y = 63$ for children vs. $\sigma_x = 47$, $\sigma_y = 56$ for parents). In addition, the heatmaps with head motion have greater vertical variance for both children and parents, suggesting that head turns from side to side may not be accompanied by gaze shifts and that participants turned their head to change what they attended. Meanwhile, head tilts upward and downward were likely accompanied with eye movements, suggesting that participants switched their attention through both head and eye movements.

Holding activities. We next examined two specific contexts involving holding objects: (1) video frames in which the individual was not holding an object, and (2) video frames in which the individual was holding an object for a sustained period of time (operationally defined as between 2 and 8 seconds). Heatmaps of eye-head alignment for these two cases are shown in Figure 2(c). When not holding, gaze data are somewhat spread out around the center in both child and

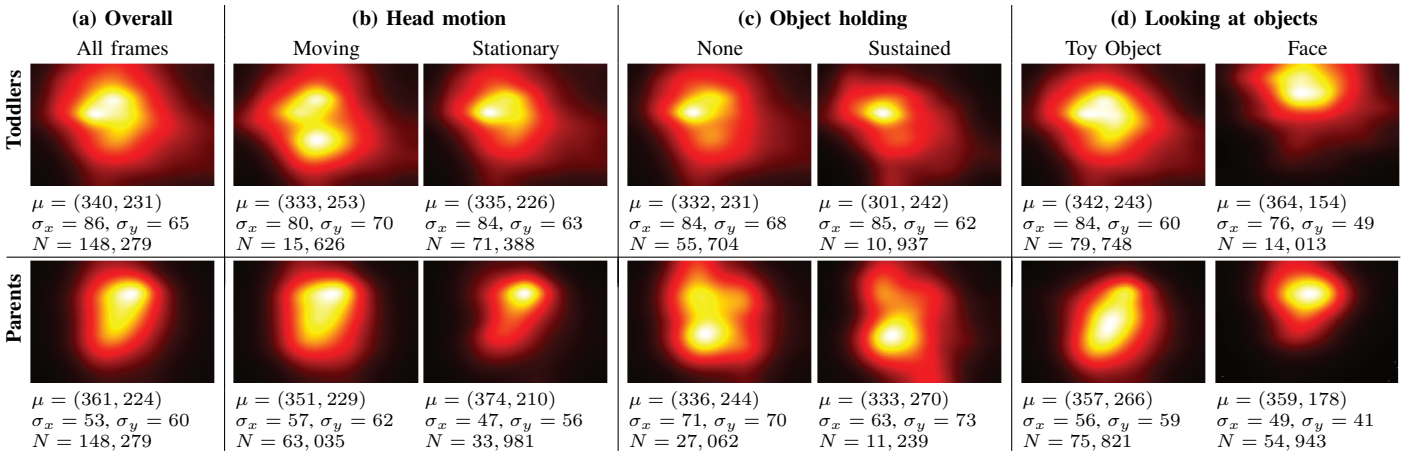


Fig. 2. Heat maps showing spatial distributions of eye gaze within first-person images, for toddlers (top) and parents (bottom), in different contexts: (a) overall across all video frames, (b) when head is moving and stationary, (c) when individuals are not holding an object or holding it for a sustained period of time (2 to 8 seconds), (d) when gaze is on an object or a face. Also shown are robust (60% trimmed) estimates of mean and variance (in horizontal (x) and vertical (y) directions) of the gaze distributions and number of gaze data points used to generated heatmaps. (Off-diagonal covariances not shown.) Maps are 720×480.

parent views. In contrast, sustained holding created moments in which eyes are more likely to fix on a certain area in view. In summary, holding actions (especially with a certain period of time) stabilize visual attention toward the center of the visual field. More generally, holding objects may also provide a solution via the coupling of head, hand and eye so as to create a stabilized visual field without head movements.

Gazing at objects and faces. We next calculated and compared heatmaps at the moments when individuals were looking at objects with those moments when they were looking at the other person's face. As shown in Figure 2(d), when looking at the partner's face, both child and parent gazed in an upper area in their first person view ($\mu = (376, 173)$ for children and $\mu = (354, 186)$, where $(360, 240)$ is the true frame center), indicating that they looked up toward the other's face without completely orienting their head toward the face. This is likely caused by brief glances at the other's face, without moving the head, prior to immediately returning attention back to objects.

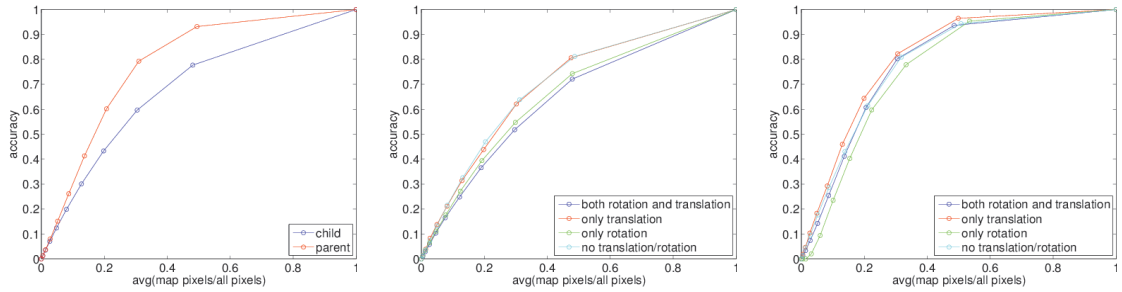
Figure 2(d) also presents the gaze distributions when looking at objects, showing an overall similar pattern with gaze distributed around the center of the visual field. Since a parent's view is much broader (see Figure 1), we can roughly define three workspaces in her view based on distance from her body: (1) a personal workspace for herself when she manipulates an object close to her body, (2) a workspace for the child when the child manipulates an object close to his body, and (3) a joint workspace midway between child and parent. Based on this definition, the parent's first-person view will typically include the personal workspace at the bottom, the joint workspace in the middle of the frame, and the child's workspace at the top. The parent gaze distribution has a vertical-shaped ellipsoid shape, suggesting that her attention is allocated (and potentially switches frequently) between all three spaces. This is likely caused by parents paying attention to both objects in their own hands as well as monitoring the objects in a child's hands.

In contrast, the child gaze heatmap is more spread out ($\sigma_x = 84$, $\sigma_y = 60$ vs. $\sigma_x = 56$, $\sigma_y = 59$ for parents), potentially because their eye-hand coordination is not fully developed. Further, because the child is shorter, the child's view is narrower than the parent's view, suggesting that they cannot switch between the workspaces without head turns, yielding a more centered heatmap. There are two possible explanations for this observation: (1) they might primarily focus on one workspace (presumably their own) when manipulating objects; or (2) they might switch their attention equally frequently among three workspaces as the parents did, but accomplishing this by moving their head and eyes together. Combining these observations with the results from the child's face gaze heatmap, brief face glances do not require a head turn to completely orient the head toward the face, while attention on objects held by the parent may be longer and require both eye and head movements. Meanwhile, parents switch their attention frequently between the child's face, objects in her own hand and objects in the child's hands, and some of these attention shifts may be executed without head turns.

B. Visual saliency

Our next goal was to measure the saliency of the visual environment from the perspective of children and adults, and then to connect gaze and attention shifts to properties of the momentary first-person salience map. Toward these goals, we employed the saliency algorithm of Itti *et al* [23], which is a biologically-motivated model that estimates a two-dimensional map of the saliency of objects in the visual environment, purely based on the properties of the visual stimulus. Briefly summarized, the model extracts low level vision features (we use motion, intensity and orientation) from a color video frame at several spatial scales (using Gaussian pyramids consisting of progressive low-pass filtering and sub-sampling of the input image). Next, each feature is computed in a center-surround structure akin to visual receptive fields, implemented as differences between a fine and a coarse scale for each feature. Finally, the visual input in the original image is represented in the form of iconic (appearance-based) topographic feature maps in which each pixel includes an estimate of its saliency.

It is important to note that while this saliency algorithm is known to have some cognitive validity, we do *not* assume that it is a perfect cognitive model. Rather, we employ this algorithm to automatically produce reasonable saliency maps, which we use to measure properties of momentary first-person views to measure how well they correlate with where



(a) (b) (c)

Fig. 3. ROC-like performance curves showing accuracy of saliency maps in predicting gaze location in first-person views. (a) Overall curves for children and adults on all image frames. (b) and (c): Curves on subsets of frames with different types of head motion, for (b) children and (c) parents.

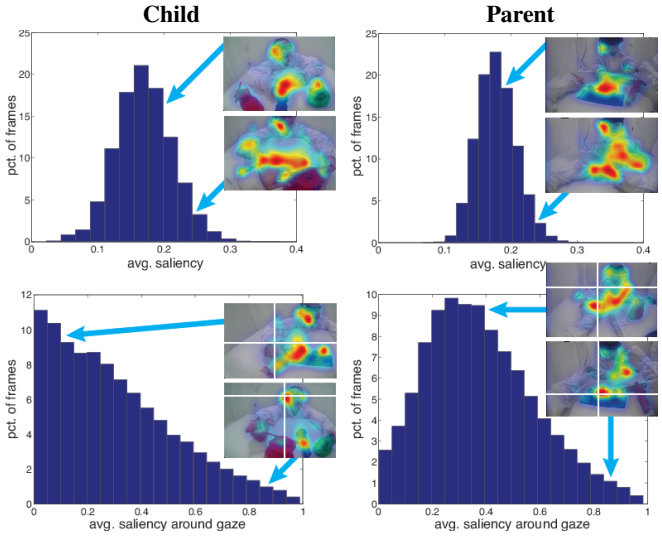


Fig. 4. Histograms of visual saliency in first person views, for children (left column) and parents (right column). *Top row:* Overall average saliency histograms, showing that distribution of saliency is about the same for the children and parents. *Bottom row:* Average saliency in a local neighborhood around the individual's gaze, showing that saliency is much more predictive of adult eye gaze position. We also show a few sample frames overlaid with saliency heat maps, and indicate which histogram bin each frame would be counted in. In the bottom row, two frames included in each plot show two representative cases, one in which the person looked at a high-saliency area, and the other in which the participant attended to a low-saliency area.

people look. In limited experimentation with other saliency algorithms including the technique of Harel et al [26], we found relatively stable results; we leave detailed comparison of saliency algorithms in first-person views to future work.

Saliency in first-person views. We generated saliency maps for a total of 296,000 frames (148,000 frames each from parent and child head cameras). An initial question is how the overall saliency of a child's view compares to that of a parent's view. The top row of Figure 4 show the distributions based on mean saliency values aggregated across all the pixels in each image frame, indicating no significant difference in overall visual saliency in child and parent views. Next, we took the tracked gaze location in each image and created a spotlight circle with a radius of 15 pixels, and calculated the mean saliency value in this local patch around each gaze point. The bottom row of Figure 4 shows that saliency values around gaze locations are significantly different between child and parent: (1) there is a larger proportion of image frames in which gaze locations are not salient at all in the child's view compared with parents view (child 12% vs. parent 3%), and (2) a large proportion of image frames in the parent's view has saliency value around 0.3 while saliency values around the child's gaze are skewed lower. This indicates that saliency may be less correlated with where children looked compared with where parents looked.

Even though our saliency model is limited to be a bottom-up approach without any top-down volitional component, our application of saliency in first person views is unique because it captures dynamic head-camera images that change rapidly due to head and body movements. Compared with saliency maps generated from static images, saliency maps of dynamic first-person views are partially created by head and hand movements. Therefore they reflect not only bottom-up saliency in visual stimuli but are also influenced by top-down bodily actions. This top-down ingredient extends the conventional definition of saliency maps and may make first-person view saliency a predictive factor of where people look. Given this view, we next attempt to use saliency maps to predict gaze.

Predicting gaze with saliency. Saliency algorithms are typically evaluated by asking a participant to look at a static image (like a photo) and then recording the series of locations at which the participant fixated over time [26]. By counting these fixations as positives and all other locations as negatives, thresholding the saliency map yields some fraction of positive locations that are correctly labeled (true positive rate) and some fraction of negative locations that are incorrectly labeled positive (false positive rate). Varying over multiple thresholds yields an ROC curve for the performance of the algorithm.

We are interested in a very different context, in which we estimate the saliency of a frame from first-person video data and then decide where a person is likely to gaze. Posing this as a classification task is complicated by the fact that every frame has at most one true positive or one false negative location because one either predicts the gaze location or one does not. To overcome this problem, we defined a pseudo-ROC curve for video data where we computed the fraction of frames in which gaze was correctly predicted over the number of all frames, instead of a true positive rate. Similarly, instead of a false positive rate, we computed the fraction of salient pixels over all pixels in the image, taking the average over all frames.

Figure 3(a) shows these pseudo-ROC curves, indicating that saliency is significantly better at predicting gaze for parents than for children. For example, at a threshold at which 40% of pixels are marked as candidate gaze positions, there is about a 70% chance that the true gaze is marked as a candidate for the toddler versus a nearly 85% chance for the adult. To understand this result better, we subdivided video frames

into four general classes: (1) head position is stationary, (2) head is rotating, (3) head is translating, and (4) head is both rotating and translating. Figures 3(b) and (c) show the curves for children and parents, respectively, in these cases. The child results show that saliency predicts gaze direction better when the head is not moving or is only translating. For adults, the performance across different head motions is similar, but again saliency is more predictive during translation than rotation. Taken together with the results from eye-head-hand coordination above that showed a closer coupling in children than in parents, these results indicate that the saliency model has limited success in capturing the influence of bodily actions on visual attention, which seems necessary to build better models of egocentric view saliency, especially for children.

V. Discussion and conclusion

Our study focuses on understanding the sensory-motor dynamics of visual attention in active task contexts using head-mounted eye tracking. As one of the first studies on tracking toddler eye movements in the real world, we document the role of eye, head, and hand actions in the selection and stabilization of visual attention on objects. Compared to adults, toddlers move a lot when they are interacting with objects. These large movements present toddlers with new attentional challenges relative to their less active infant selves. However, our results show that young children's attentional systems are more tied to bodily action compared with adults, and suggest that manual action (holding objects) may also provide a solution to stabilize attention on one object via the coupling of head, hand and eye. We found that head turns also play an important role in stabilizing attention. Early in human development, attention is deeply linked to whole body movements, which create large changes in the relative salience of different components of the scene. To quantify the link between visual saliency and gaze in first-person views, we applied a well-established saliency algorithm to egocentric video and found it is more predictive when the head is stationary. Since the head moves a lot in free-flowing interaction, the predictive ability of saliency is above chance but far from perfect, which is not surprising given that most saliency models are designed for stationary tasks such as picture viewing [16]. Nonetheless, this observation poses a new challenge to build saliency algorithms for egocentric views, and our quantitative results on eye, head and hand coordination provide useful empirical patterns toward this goal.

In summary, we have presented a first step towards understanding visual attention of toddlers in naturalistic tasks using head-mounted eye and motion tracking. Our results suggest that for toddlers, effective visual attention requires stabilizing the head and aligning the head and eyes, behaviors fostered by actions such as reaching for and holding objects. The conjecture is that these bodily alignments also align the internal spatial representations of the attended location, stabilizing visual attention and making that attention more effective for development and learning. In future work, this finding may also help build models of visual attention in egocentric views.

Acknowledgment

This work was supported by grants from the IU Offices of the Vice President and Vice Provost for Research through the Faculty Research Support Program and from the National Science Foundation (BCS-0924248).

References

[1] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychol. Bull.*, vol. 124, no. 3, pp. 372–422, 1998.

[2] A. Seitz, D. Kim, and T. Watanabe, "Rewards evoke learning of unconsciously processed visual stimuli in adult humans," *Neuron*, vol. 61, pp. 700–707, 2009.

[3] M. G. Shuler and M. F. Bear, "Reward timing in the primary visual cortex," *Science*, vol. 311, pp. 1606–1609, 2006.

[4] L. B. Smith, C. Yu, and A. F. Pereira, "Not your mother's view: the dynamics of toddler visual experience," *Developmental Science*, vol. 14, no. 1, pp. 9–17, 2011.

[5] C. Yu, L. B. Smith, H. Shen, A. F. Pereira, and T. Smith, "Active information selection: Visual attention through the hands," *IEEE Trans. on Auton. Ment. Devel.*, vol. 1, no. 2, pp. 141–151, 2009.

[6] J. Franchak, K. Kretch, K. Soska, and K. Adolph, "Head-mounted eye tracking: A new method to describe infant looking," *Child development*, vol. 82, no. 6, pp. 1738–1750, 2011.

[7] M. Posner, "Orienting of attention," *Quarterly Journal of Experimental Psychology*, vol. 32, no. 1, pp. 3–25, 1980.

[8] S. P. Vercera and S. J. Luck, "Attention," in *Encyclopedia of the human brain*, vol. 1. Academic Press, 2002, pp. 269–284.

[9] M. Shepherd, J. Findlay, and R. Hockey, "The relationship between eye movements and spatial attention," *Quarterly Journal of Experimental Psychology*, vol. 38A, pp. 475–491, 1986.

[10] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth, "Variability of eye movements when viewing dynamic natural scenes," *J. Vis.*, vol. 10, no. 10, pp. 1–17, 2010.

[11] M.-H. Grosbras, A. R. Laird, and T. Paus, "Cortical regions involved in eye movements, shifts of attention, and gaze perception," *Human Brain Mapping*, vol. 25, no. 1, pp. 140–154, 2005.

[12] G. Rizzolatti, L. Riggio, I. Dascola, and C. Umiltá, "Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention," *Neuropsychologia*, vol. 25, 1987.

[13] C. L. Colby and M. E. Goldberg, "Space and attention in parietal cortex," *Annu. Rev. Neurosci.*, vol. 22, pp. 319–349, 1999.

[14] D. J. Hagler, L. Riecke, and M. I. Sereno, "Pointing and saccades rely on common parietal and superior frontal visuospatial maps," *Neuroimage*, vol. 35, pp. 1562–1577, 2007.

[15] E. Knudsen, "Fundamental components of attention," *Annual Review of Neuroscience*, vol. 30, pp. 57 – 78, 2007.

[16] B. Tatler, M. Hayhoe, M. Land, and D. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *J. Vis.*, vol. 11, no. 5, pp. 1–23, 2011.

[17] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends in Cognitive Sciences*, vol. 9, no. 4, pp. 188–194, 2005.

[18] F. Raudies, R. Gilmore, K. Kretch, J. Franchak, and K. Adolph, "Understanding the development of motion processing by characterizing optic flow experienced by infants and their mothers," in *ICDL*, 2012.

[19] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *CVPR*, 2012.

[20] A. Fathi, X. Ren, and J. Rehg, "Learning to recognize objects in egocentric activities," in *CVPR*, 2011.

[21] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *CVPR*, 2012.

[22] A. Fathi, Y. Li, and J. Rehg, "Learning to recognize daily actions using gaze," in *ECCV*, 2012.

[23] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *T. PAMI*, vol. 20, no. 11, 1998.

[24] J. Pelz, M. Hayhoe, and R. Loeber, "The coordination of eye, head, and hand movements in a natural task," *Experimental Brain Research*, vol. 139, no. 3, pp. 266–277, 2001.

[25] P. H. C. Eilers and J. J. Goeman, "Enhancing scatterplots with smoothed densities," *Bioinformatics*, vol. 20, no. 5, pp. 623 – 628, 2004.

[26] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2007, pp. 545–552.