# Supporting Information

## Ullman et al. 10.1073/pnas.1207690109

### SI Materials and Methods

Training and test data consisted of sets of video sequences, listed below. Example video clips are included as Movies S1–S7. The data sets were selected to cover a broad range of stimuli. Different stimuli supply useful information to different stages of the learning process; however, the relevant stimuli are extracted automatically by the algorithm and do not need to be presented in a specific order. For example, mover events are detected almost exclusively in videos containing manipulating hands. All videos were black and white, taken with a stationary video camera. Videos were scaled to make hand width roughly 40 pixels, selected according to the visual acuity of infants at approximately 3–6 mo of age, observing the hand from a distance of up to 1.5–2 m (1).

1. *Movers:* Combination of hands manipulating objects and autonomously moving objects (e.g., rolling balls). Four video sequences, total of 22,545 frames (~15 min), 360 × 288 pixels. Each video shows one of three actors sitting behind a table and manipulating objects on the table.
2. *Manipulating hands:* Actors engaged in object manipulation, picking up, moving, and placing objects. Eight video sequences, total of 9,122 frames (~6 min), 350 × 400 pixels. Each video shows one of four actors facing the camera on one of two backgrounds, standing behind a table and manipulating objects on the table, one hand at a time.
3. *Freely moving hands:* Eight video sequences, total of 11,823 frames (~8 min), 436 × 336 pixels. Each video shows one of four different actors facing the camera on one of two backgrounds. Actors move their hands around, one hand at a time (the other at the side of the body), occasionally performing one of six different gestures.
4. *Walk and manipulate:* People walking and occasionally manipulating objects. One video sequence, ~15,000 frames (~10 min), 360 × 288 pixels. Video shows several people passing by a table and occasionally putting objects on the table or picking them up.
5. *Walking:* People walking, without object manipulations. One video sequence, 4,530 frames (~3 min), 702 × 576 pixels. Video shows two actors walking back and forth.
6. *Own hands:* Moving hands from a first-person perspective. Two video sequences, total of 6,388 frames (~4 min), 144 × 112 pixels. Each video shows the hands of an actor moving around. Video is taken with the camera near the actor's head.
7. *Gaze:* Actors moving objects on a table. Eight video sequences, total of 34,631 frames (~23 min), 540 × 432 pixels. Each video shows a different actor moving objects between different locations on a table. Eight spots were marked on the table at different locations. Six objects were placed on six of these spots. Actors were instructed to pick up one object at a time and put it at an empty spot. No instructions were given concerning gaze. Each video contains approximately 50 object moves, consisting of picking up and placing at a different location.

### SI Results

**Alternative Cues.** In the model, the initial learning of hand detection uses active motion, or mover events, as a cue for potential hand regions in the image. We compared this learning with a number of alternative cues, including saliency, object-containing regions, and information-based and motion-based cues. The comparisons served two goals: first, to compare general object learning methods, which are not specific to hands, with methods that rely on domain-specific cues, biased toward hand stimuli; second, to compare the mover-based cue with two natural alternatives, general motion cues and the use of own-hands.

We compared hand-candidate regions produced by different cues with annotated ground-truth, by counting the fraction of successful hits (center-to-center distance between the candidate and true hand up to 30 pixels) of the top 100 and the top 2,500 hand candidates suggested by each cue. (The movers did not produce a score; we selected randomly 2,500 out of the total 3,567 candidates.) We also used the different cues to train an object detector (2, 3) (same for all cues), using 2,500 patches of 90 × 90 pixels taken around the top scoring hand candidates selected by each cue. Nonclass examples were randomly chosen 90 × 90 patches adjacent to the class patches in the same original images. Hand candidates were extracted from the *Walk and manipulate* dataset. The resulting detectors were tested on both the *Manipulating hands* and *Freely moving hands* datasets.

**Saliency.** The main alternatives of interest rely on general rather than hand-specific cues. The saliency-based alternative assumes that hand-containing regions may attract attention on the basis of general salience cues, leading to a biased processing of hand-regions. We applied a state-of-the-art saliency detector (4) to each frame of the training video clips and extracted the 10 most salient image locations at each frame and their saliency score. Only a small fraction (0.5%; Table S1) of the top 2,500 salient regions contained hands, which was insufficient for training a hand detector.

**Generic Object Detection.** A recent computational alternative to general saliency is a scheme for locating image regions that are likely to contain objects of interest [termed "generic objectness" measure (5)]. This measure was shown to out-perform saliency models in several comparative tests. Only a small fraction (0.2%; Table S1) of the top 2,500 detected regions contained hands, which was insufficient for training a hand detector.

**Informative Fragments.** We selected informative image regions according to mutual information criteria, which often extract a high fraction of class features for natural class (6). The algorithm extracts image features that are characteristic to a class. We used regions containing people as class examples and regions without people as nonclass. We examined whether hand regions are selected among the informative person-specific regions. A small fraction (less than 2%; Table S1) of the top 2,500 detected regions contained hands, which resulted in a poorly performing hand detector (Fig. S1).

**Image Motion.** We compared the active-motion cues with the use of general motion. This is close to the mover-based method, but using general image motion without distinguishing active from other forms of motion. To test different types of motion, we used two video sequences for training, *Walk and manipulate* and *Walking*.

For general motion cues, we calculated the optical flow (7) for each video frame and selected points with high optical flow magnitude. In the mixture of the two video sequences (taking an equal number of hand candidates from each video) the fraction of hand-containing regions was 16.8% in the top 2,500 motion regions (Table S1). Detection performance of a hand detector learned from these examples was inferior to that of a similar detector trained on movers-based examples from the same videos (e.g., 27% vs. 91% correct detection at 2% recall rate; Fig. S1). Markedly, no mover events were detected in the *Walking* video sequence (which contains no object manipulations). For general motion, detection results for free-moving hands were close to manipulating hands, in contrast with the mover-based detection whereby manipulating hands produce highly improved results.

**Learning from Own Hands.** A possible source of information for learning about hands can be obtained by moving and observing one's own hands. Both behavioral (8) and physiological (9) evidence support the use and representation of "own hands," but their possible role in developing hand detection remains unclear. In our testing, own-hand images were obtained from two adult subjects using video cameras placed close to the subject's head (example images in Fig. S2). The training images allow us to evaluate the limitations of own-hand images under favorable training conditions (good imaging conditions and a broad range of hand configurations); realistic images obtained from an infant's perspective may be less informative and will be interesting to analyze in future studies. Total number of examples used for training was similar to the movers training. Using longer training sequences of similar data had minor effect on performance. As in other methods, testing was done on the *Manipulating hands* and *Freely-moving hands* datasets.

For detector training, the *Own hands* training videos were used to train a hand detector using positive and negative examples. Optical flow (7) was used to extract image patches containing large moving parts. These image patches mostly contain hands (Fig. S2) and provided positive class image examples. Negative nonclass examples were patches extracted from nonperson images. The same detector used for the main mover-based scheme (2, 3) was trained on these examples.

Fig. S3 shows an average precision-recall graph for the two hand detectors when applied to the *Manipulating hands* and *Freely moving hands* datasets. The results show that, in contrast to the mover-based detection, the own-hands detector did not generalize to the *Manipulating hands* dataset (maximal precision less than 5%). For *Freely moving hands*, in the low-recall (2%), high-precision range (which is used for subsequent training), own-hands reached approximately 15% precision compared with 97% of the mover-based. It is interesting to note that data from infants' behavior indicates that their looking time is particularly high for hands engaged in object manipulation (10, 11). This finding is more consistent with computational results obtained from mover-based training than with own-hands training, probably because generalization is more straightforward.

1. Teller DY, McDonald MA, Preston K, Sebris SL, Dobson V (1986) Assessment of visual acuity in infants and children: The acuity card procedure. *Dev Med Child Neurol* 28:779–789.
2. Crandall D, Felzenszwalb P, Huttenlocher D (2005) Spatial priors for part-based recognition using statistical models. *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), pp 10–17, 10.1109/CVPR.2005.329.
3. Karlinsky L, Dinerstein M, Harari D, Ullman S (2010) The chains model for detecting parts by their context. *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), pp 25–32, 10.1109/CVPR.2010.5540232.
4. Walther DB, Koch C (2006) Modeling attention to salient proto-objects. *Neural Netw* 19:1395–1407.
5. Alexe B, Deselaers T, Ferrari V (2010) What is an object? *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), pp 73–80, 10.1109/CVPR.2010.5540226.
6. Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. *Nat Neurosci* 5:682–687.
7. Black MJ, Anandan P (1996) The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Comput Vis Image Underst* 63:75–104.
8. Sommerville JA, Woodward AL, Needham A (2005) Action experience alters 3-month-old infants' perception of others' actions. *Cognition* 96:B1–B11.
9. Caggiano V, et al. (2011) View-based encoding of actions in mirror neurons of area f5 in macaque premotor cortex. *Curr Biol* 21:144–148.
10. Aslin RN (2009) How infants view natural scenes gathered from a head-mounted camera. *Optom Vis Sci* 86:561–565.
11. Frank M, Vul E, Saxe R (2012) Measuring the development of social attention using free-viewing. *Infancy* 17:355–375.

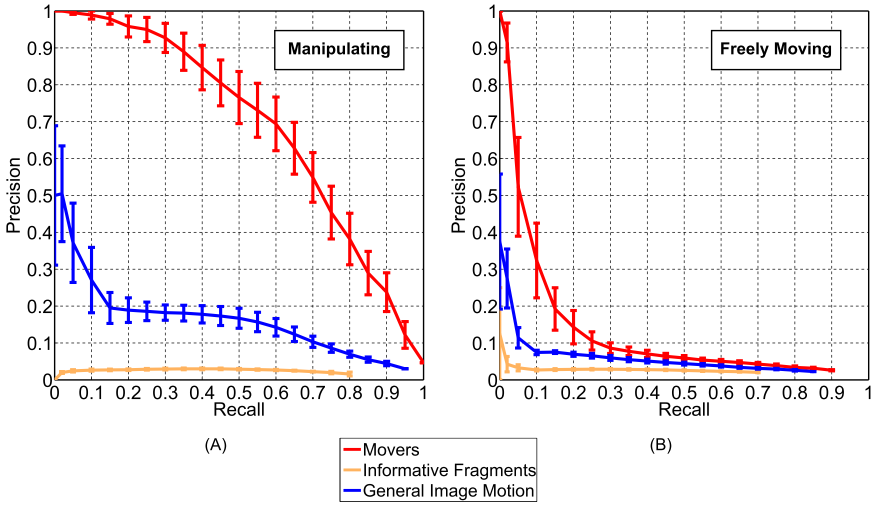


**Fig. S1.** Comparison of hand detectors trained by different cues. Precision-recall curves show mean and SE over all video sequences of (*A*) the *Manipulating hands* dataset, and (*B*) the *Freely moving hands* dataset. Red, training by mover events; blue, training by general motion; yellow, training by information maximization. Abscissa: recall rate; ordinate: precision.

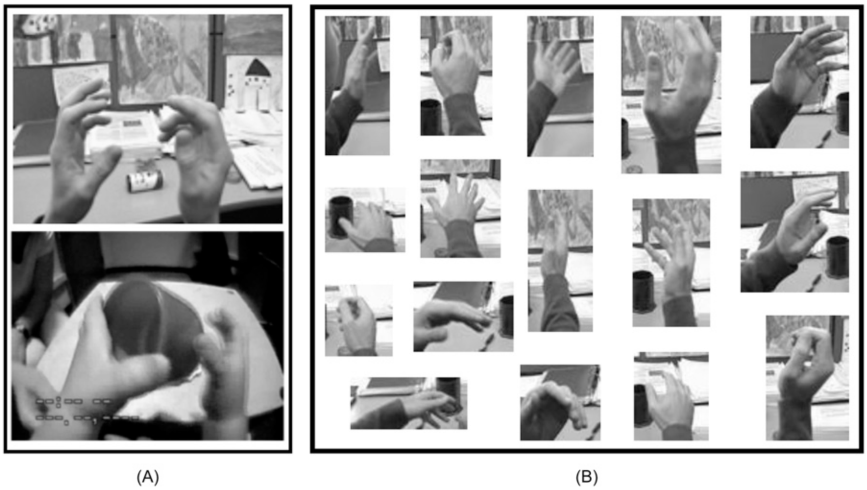


**Fig. S2.** Data used for training the own-hands detector. (*A*) *Upper*: Example frame from the training video. *Lower*: Video frame taken from an infant head-mounted camera [adapted from Yoshida and Smith (1)]. (*B*) Training examples extracted automatically from the *Own hands* training video.

[1] Yoshida H, Smith LB (2008) What's in view for toddlers? Using a head camera to study visual experience. *Infancy* 13:229–248.

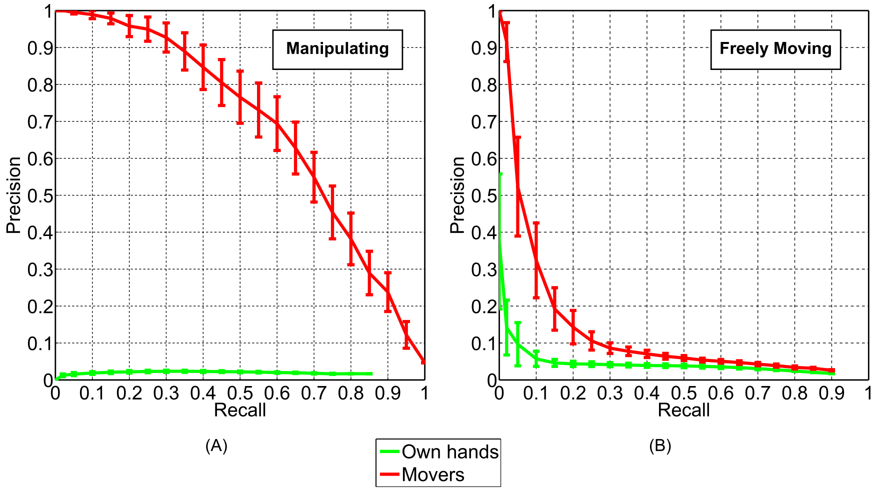


**Fig. S3.** Comparison of hand detectors trained by mover events and by own-hands. Precision-recall curves show mean and SE over all video sequences of (*A*) the *Manipulating hands* dataset, and (*B*) the *Freely moving hands* dataset. Red, training by mover events; green, training from own hands. Abscissa: recall rate; ordinate: precision.

**Table S1. Hand extraction by different cues**

| Cue | Precision for 2,500 best candidates (%) | Precision for 100 best candidates (%) |
|---|---|---|
| Saliency (4) | 0.5 | 0.0 |
| Generic object detector (5) | 0.2 | 0.0 |
| Informative fragments (6) | 1.8 | 3.0 |
| Image motion | 16.8 | 16.0 |
| Movers | 64.7 | |

Hand detectors were trained on the highest-scoring 2,500 hand candidates. Also shown are the top 100 candidates.

**Movie S1.** Examples from *Movers* dataset: combination of hands manipulating objects and autonomously moving objects.

Movie S1

**Movie S2.** Examples from *Manipulating hands* dataset: hands engaged in object manipulation, picking up, moving and placing object.

Movie S2

**Movie S3.** Examples from *Freely moving hands* dataset: hands at a broad range of natural poses, showing entire upper body.

Movie S3

**Movie S4.** Examples from *Walk and manipulate* dataset: people walking and occasionally manipulating objects.

Movie S4

**Movie S5.** Examples from *Walking* dataset: people walking, without object manipulations.

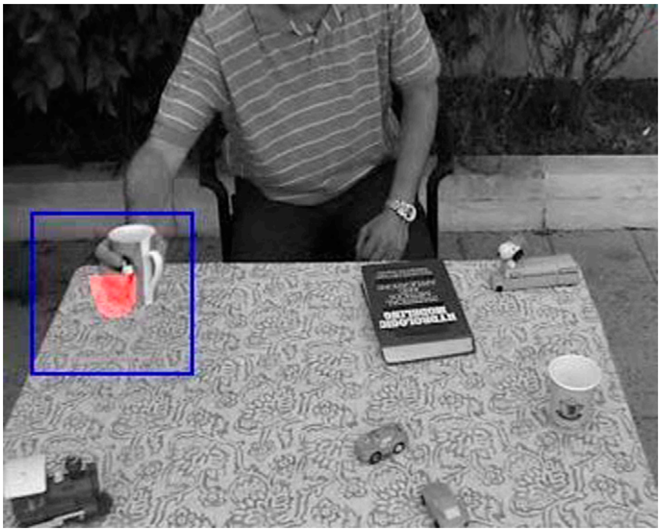
Movie S5

**Movie S6.** Examples from *Own hands* dataset: moving hands from a first-person perspective.

Movie S6

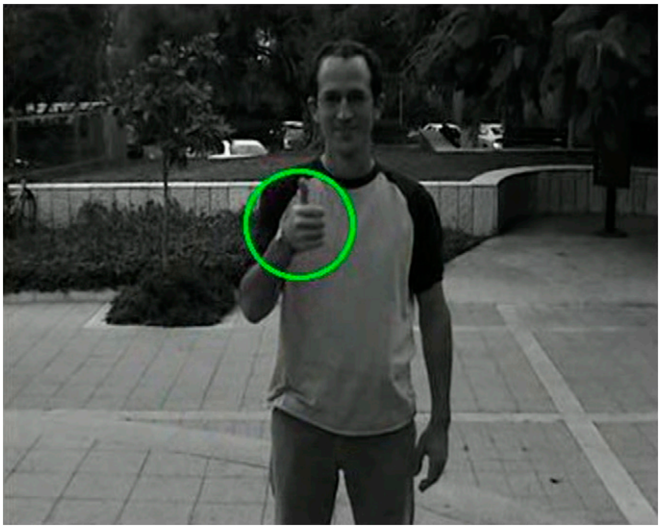
**Movie S7.** Examples from *Gaze* dataset: actors moving objects on a table.

Movie S7



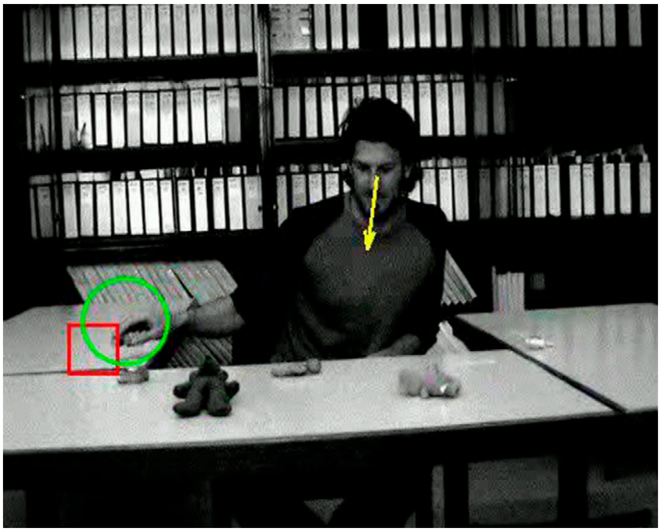
**Movie S8.** Example of mover detection results: detected mover events and tracked movers. Green, last outgoing pixels from cell of event; red, moving pixels of the tracked mover (restricted to a 30 × 30 region); blue rectangle, the 90 × 90 image patch extracted as a candidate hand.

Movie S8



**Movie S9.** Example of hand detection results: hand detections of the final hand detector. Green circle, detected hands; yellow circle, tracked hands.

Movie S9



**Movie S10.** Example of combined results: mover events, detected hands, and predicted gaze direction. Red square, mover events; green circle, detected and tracked hands; yellow arrow, predicted gaze direction.

Movie S10