SEE COMMENTARY

From simple innate biases to complex visual concepts

Shimon Ullman[1,2], Daniel Harari[1], and Nimrod Dorfman[1]

Department of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 76100, Israel

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved September 4, 2012 (received for review May 16, 2012)

Early in development, infants learn to solve visual problems that are highly challenging for current computational methods. We present a model that deals with two fundamental problems in which the gap between computational difficulty and infant learning is particularly striking: learning to recognize hands and learning to recognize gaze direction. The model is shown a stream of natural videos and learns without any supervision to detect human hands by appearance and by context, as well as direction of gaze, in complex natural scenes. The algorithm is guided by an empirically motivated innate mechanism—the detection of "mover" events in dynamic images, which are the events of a moving image region causing a stationary region to move or change after contact. Mover events provide an internal teaching signal, which is shown to be more effective than alternative cues and sufficient for the efficient acquisition of hand and gaze representations. The implications go beyond the specific tasks, by showing how domain-specific "proto concepts" can guide the system to acquire meaningful concepts, which are significant to the observer but statistically inconspicuous in the sensory input.

cognitive development | hand detection | unsupervised learning | visual cognition

A basic question in cognitive development is how we learn to understand the world on the basis of sensory perception and active exploration. Already in their first months of life, infants rapidly learn to recognize complex objects and events in their visual input (1–3). Probabilistic learning models, as well as connectionist and dynamical models, have been developed in recent years as powerful tools for extracting the unobserved causes of sensory signals (4–6). Some of these models can efficiently discover significant statistical regularities in the observed signals, which may be subtle and of high order, and use them to construct world models and guide behavior (7–10). However, even powerful statistical models have inherent difficulties with natural cognitive concepts, which depend not only on statistical regularities in the sensory input but also on their significance and meaning to the observer. For example, in learning to understand actions and goals, an important part is identifying the agents' hands, their configuration, and their interactions with objects (1–3). This is an example in which significant and meaningful features can be nonsalient and highly variable and therefore difficult to learn. Our testing shows that current computational methods for general object detection (11–13) applied to large training data do not result by themselves in automatically learning about hands. In contrast, detecting hands (14), paying attention to what they are doing (15, 16), and using them to make inferences and predictions (1–3, 17) are natural for humans and appear early in development. How is it possible for infants to acquire such concepts in early development?

A large body of developmental studies has suggested that the human cognitive system is equipped through evolution with basic innate structures that facilitate the acquisition of meaningful concepts and categories (9, 15, 18–21). These are likely to be not developed concepts, but some form of simpler "proto concepts," which serve as anchor points and initial directions for the subsequent development of more mature concepts. Major questions that remain open are the nature of such innate concepts and how they can guide the subsequent development of mature concepts.

Here we show in a model that the incorporation of a plausible innate or early acquired bias, based on cognitive and perceptual findings, to detect "mover" events (Fig. 1) leads to the automatic acquisition of increasingly complex concepts and capabilities, which do not emerge without domain-specific biases. After exposure to video sequences containing people performing everyday actions, and without supervision, the model develops the capacity to locate hands in complex configurations by their appearance and by surrounding context (Fig. 2 *A–C*) and to detect direction of gaze (Fig. 3 *D* and *E*).

Hands are frequently engaged in motion, and their motion could provide a useful cue for acquiring hand concepts. Infants are known to have mechanisms for detecting motion, separating moving regions from a stationary background, and tracking a moving region (22, 23). However, our simulations showed that general motion cues on their own are unlikely to provide a sufficiently specific cue for hand learning: the extraction of moving regions from test video sequences can yield a low proportion of hand images, which provides only a weak support for extracting the class of hands (*SI Results* and Fig. S1). Infants are also sensitive, however, to specific types of motion, including launching, active (causing other objects to move), self-propelled, or passive (24–27). On the basis of these findings, we introduced in the model the detection of active motion, which we call "mover" event, defined as the event of a moving image region causing a stationary region to move or change after contact (*Methods*, Fig. 1 *A–C*, and Movie S8). Mover detection is simple and primitive, based directly on image motion without requiring object detection or region segmentation.

Results

Our model detects mover events and uses them to train a hand classifier. Training and testing of the entire model used a total of 30 video sequences (approximately 65 min) showing people moving about, moving their hands, manipulating objects, and some moving objects (*SI Materials and Methods* and Movies S1–S7). First we use the detected movers (Fig. 1 *D–F*) to train a standard classifier (28). Because hands are highly represented in the mover-tagged regions (approximately 65%), after this learning phase, hands are classified with high precision (97%) but with low recall rate (2%) (Fig. 2*D*). (Precision/recall measures classification performance; the classifier's output contains 2% of all hands in the input, with 97% accuracy.) Recall is initially low because detection is limited to specific hand configurations extracted as movers, typically engaged in object grasping (Fig. 1 *D* and *E*). Recall rate rapidly increases in subsequent iterations of the learning algorithm based on two mechanisms: tracking and body context (Fig. 2 *D* and *E*). Detected hands are tracked over a short period, and additional hand images are extracted during

Author contributions: S.U., D.H., and N.D. designed research; D.H. and N.D. performed research; D.H. and N.D. analyzed data; and S.U., D.H., and N.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

See Commentary on page 17736.

[1]S.U., D.H., and N.D. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: shimon.ullman@weizmann.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1207690109/-/DCSupplemental.

PSYCHOLOGICAL AND COGNITIVE SCIENCES
COMPUTER SCIENCES

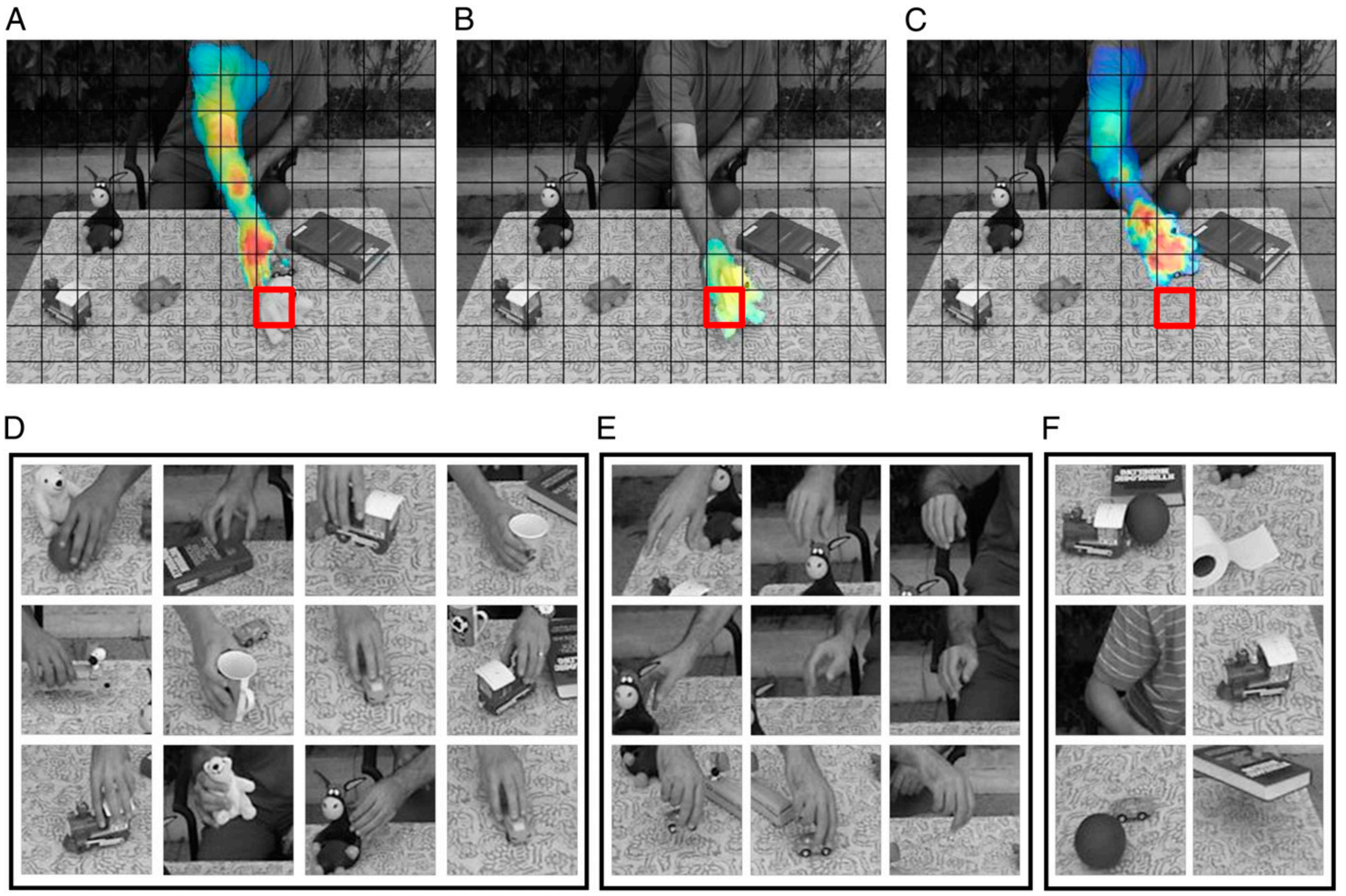


Fig. 1. Mover events. *Upper*: Mover event detected in the red cell: motion (*A*) flows into the cell, (*B*) stays briefly in the cell, and then (*C*) leaves the cell, changing its appearance. Motion is shown in color, where warmer colors indicate faster motion. *Lower*: Examples of detected movers: (*D*) detected hands, (*E*) appearance variations added by tracking (each row is taken from one track), and (*F*) nonhand detections.

the tracking period. Context is useful for detecting hands on the basis of surrounding body parts in ambiguous images. Developmental evidence indicates that infants associate hands with other body parts at approximately 6–9 mo (29) [possibly earlier with faces (30, 31)]. We therefore included in the model an existing algorithm that uses surrounding body parts, including the face, for hand detection (32).

Because hands are detected on the basis of either appearance or body context, we found that the two detection methods cooperate to extend the range of appearances and poses used by the algorithm. The context-based detection successfully recognizes hands with novel appearances, provided that the pose is already known ("pose" here is the configuration of context features, on the shoulders, arms, etc.). The newly learned appearances lead in turn to the learning of additional poses. The learning was applied to the input videos in several iterations (*Methods*). The results show that appearance-based and context-based recognition guide each other to boost recognition performance (Fig. 2 and Movie S9). Performance improves rapidly during the first three iterations, approaching the performance of fully supervised training on the same data.

Direction of Gaze. Finally, we propose that detected mover events provide accurate teaching cues in the acquisition of another intriguing capacity in early visual perception—detecting and following another person's gaze on the basis of head orientation, and later, eyes direction (33–35). This skill, which begins to develop at approximately 3–6 mo, plays an important role in the development of communication and language (36). It remains unclear how this learning might be accomplished, because cues for direction of gaze (head and eyes) can be subtle and difficult to extract and use (37).

Infants' looking is attracted by other people's hands engaged in object manipulation (15, 16), and they often shift their gaze from face to a manipulating hand (31). People often look at objects they manipulate, especially slightly before and during initial object contact. Our algorithm therefore uses mover events as an internal teaching signal for learning direction of gaze. It detects presumed object contacts by detecting mover events, extracts a face image at the onset of each mover event, and learns, by standard classification techniques, to associate the face image with the 2D direction to the contact event (Fig. 3 *A–C*). Consistent with developmental evidence (38, 39), we assume that initial face detection is present before gaze learning, and locate the face with an available face detector. The resulting classifier estimates gaze direction in novel images of new persons with accuracy approaching adult human performance under similar conditions (Fig. 3 *D* and *E* and Movie S10).

Alternative Cues. The algorithm focuses on motion-based cues, but additional visual mechanisms [e.g., biomechanical motion (40)] as well as nonvisual sensory motor cues (41, 42), supplied in part by the mirroring system (43), may also play a role in the learning process. In particular, a possible contribution can come from observing one's own hands in motion. Our testing shows, however, that using own-hands images is not as effective as using mover instances in detecting hands in general and hands engaged in object manipulation in particular (Figs. S2 and S3). In

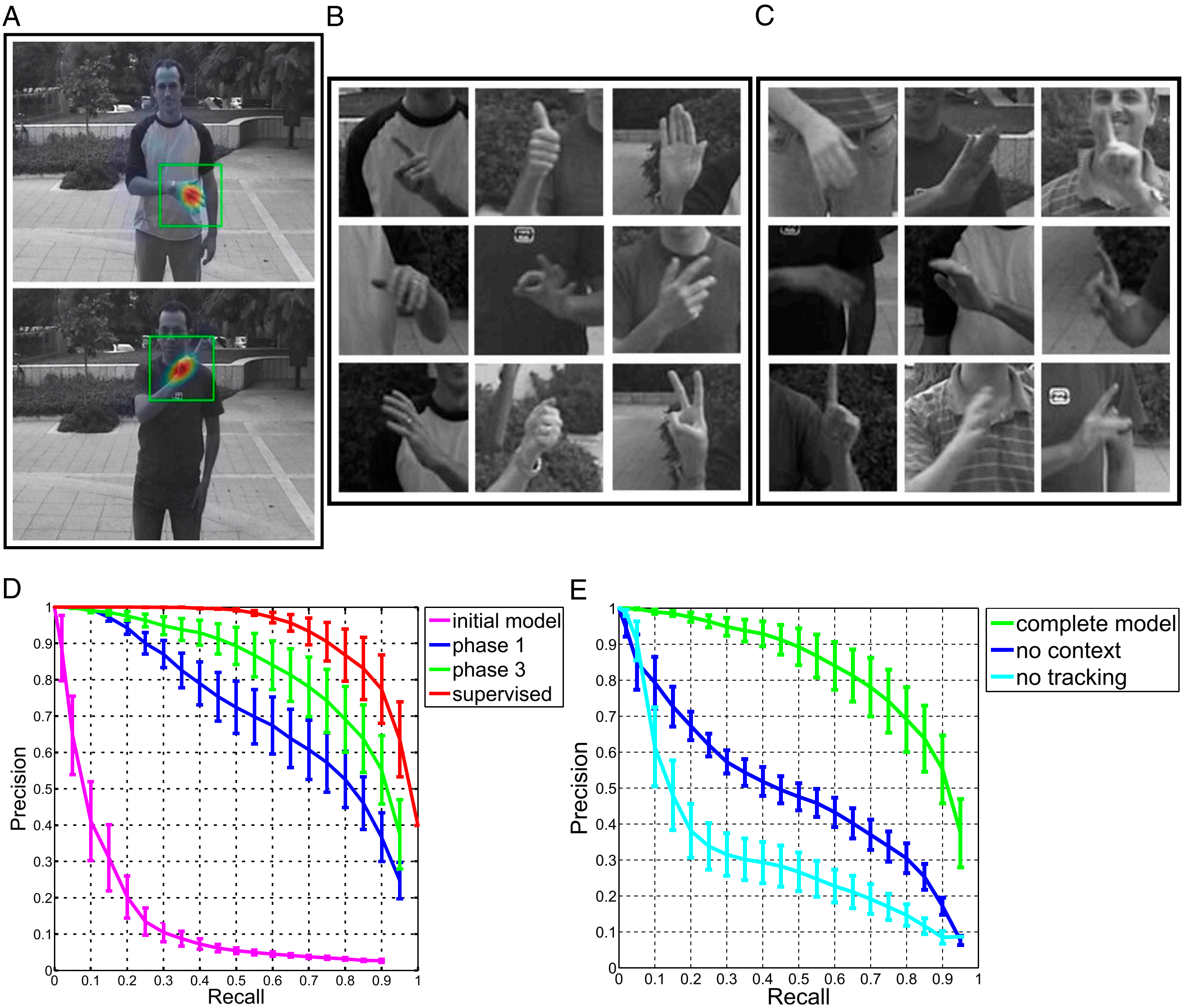


Fig. 2. Hand detection. *Upper*: Hand detection in test videos using the final detector. (*A*) Green square shows detected hands. Warm colors indicate high detection score. (*B* and *C*) Examples of hands detected by appearance (*B*) and by context (*C*). *Lower*: Hand detection performance. (*D*) Detector trained from mover events only (magenta), and after one (blue) and three (green) iterations of cotraining by appearance and context. The top curve (red) shows performance obtained with full (manually) supervised learning. (*E*) Results after three iterations of training as before (green), without using context (blue) and without tracking (cyan). Error bars indicate SE across eight folds of cross-validation. Abscissa: Recall rate (fraction of detected hands out of all of the hands in the data); ordinate: precision (fraction of hands out of all detections).

addition to "own hands," we tested a number of other potential cues in terms of their ability to provide an internal teaching signal for learning about hands (*SI Results*). First, we tested the visual saliency of hand-containing regions, using a state-of-the-art visual saliency model (13). Second, we applied a recent computational method for locating image regions that are likely to contain objects of interest (11). Finally, we applied an information-based method to extract image features that are typical of person images. Formally, these are image features that have high mutual information with respect to images of a person (12). None of these methods provided a sufficiently reliable cue for training our hand detection scheme. For example, the fraction of correct hand images produced by these three methods was <2% (out of 2,500 extracted image regions by each method), compared with 65% provided by mover events (Fig. S1 and Table S1).

Discussion

Using natural dynamic visual input and without external supervision, the algorithm learns to detect hands across a broad range of appearances and poses and to extract direction of gaze. The acquisition of these concepts is based in the model on an innate or early-learned detection of mover events, supported by infants' sensitivity to launching and active motion (24, 25), and evidence of associating hands with causing objects to move (3, 16). Future studies may resolve whether the sensitivity to mover events is based on some innate capacity or learned early in development, on the basis of an even simpler capacity. The algorithm also uses in an innate manner the spatiotemporal continuity in tracking and an association made by infants between hands and face features (30, 31).

Learning body-parts detection and gaze direction are two capacities in which the gap between computational difficulty and

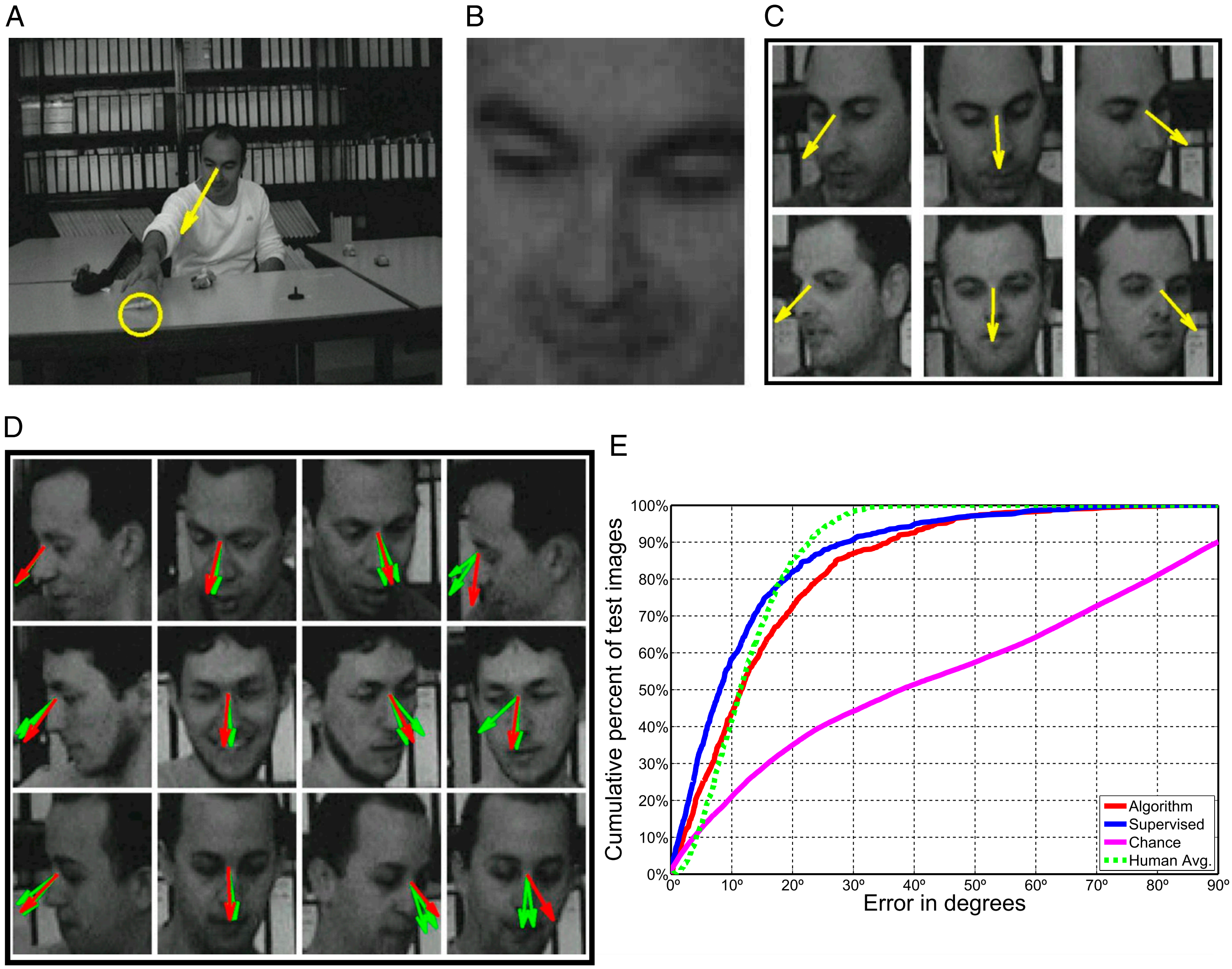


Fig. 3. Gaze direction (gaze direction recovered within the image plane). *Upper*: Training. (*A*) Detected mover event. Yellow circle marks location of event, providing teaching signals for gaze direction (yellow arrow). (*B*) Face image region used for gaze learning. (*C*) Examples of face images with automatically estimated gaze direction, used for training. *Lower*: Results. (*D*) Predicted direction of gaze: results of algorithm (red arrows) and two human observers (green arrows). (*E*) Performance of gaze detectors: detector trained on mover events (red), detector trained with full manual supervision (blue), chance level (magenta), and human performance (dashed green). Chance level was computed using all possible draws from the distribution of gaze directions in the test images. Human performance shows the average performance of two human observers. Abscissa: maximum error in degrees; ordinate: cumulative percent of images.

infant learning is particularly striking (37, 44). This model is a demonstration of learning to detect hands and gaze direction in an unsupervised manner, in natural images. The learning succeeds because the internal teaching signals, produced by the algorithm based on mover events, identify candidates of meaningful features and make them easily learnable, even when they are nonsalient and highly variable. The results of the model and computational tests complement and extend prior empirical studies: the mechanisms used by the model have been discovered and studied empirically, and the present study demonstrates that they supply surprisingly effective teaching signals for hand detection and gaze direction.

The model relies on mechanisms that seem to exist in early development, in particular, initial forms of motion analysis and face detection. It also gives rise to predictions that can be tested empirically, for example by testing for selective adaptation in infants to mover events, testing whether manipulating hands are the first hand stimuli to be acquired, and possibly preferential looking at faces after looking at mover events. Physiologically, the model suggests that if monkeys are shown during development objects being moved consistently by an artificial mover device, representations of the artificial mover may become incorporated in hand-specific brain regions, possibly in the mirror system (45). Similarly, sufficient exposure to mover events by an artificial mover is predicted to promote hand-like expectations in infants (1, 3).

On a general level, the results demonstrate a useful combination of learning and innate mechanisms: a meaningful complex concept may be neither learned on its own nor innate. Instead, a simple domain-specific internal teaching signal can guide the learning system along a path, which leads to the gradual acquisition of complex concepts that are difficult to learn and do not emerge automatically otherwise.

Materials and Methods

Data. Training and testing of the entire model used a total of 30 video sequences (approximately 65 min) showing people moving and manipulating objects (more details in *SI Materials and Methods*).

Mover Detection. To detect mover events, each video frame is divided into cells of 30 × 30 pixels. (With fixed cell size, changes in hand size of 4:1

produced similar mover detections.) A mover event is detected in cell C whenever the following sequence occurs. (*i*) C is initially stable for at least 2 s. (*ii*) A sufficient amount of visual motion entering C followed by visual motion flowing out of C within 1 s. (*iii*) C becomes stable again for at least 2 s. (*iv*) The appearance of C differs from its previous stable appearance. (Detection is not sensitive to the choice of the timing parameters). Mover detection therefore requires local motion and change detection, which are computed as follows. Motion is computed using a standard optical flow algorithm (46), which is similar to biological models (47). The computation requires only the local flow of individual pixels between adjacent cells, without delineating objects or coherent regions, and without any longer-term motion tracking. To define stability and change, we filter out cells undergoing brief transience in appearance. For transience detection, the appearance A_t of a cell at time t (its 30 × 30-pixel values) is compared with its smoothed version $\overline{A}_t$ obtained by exponential moving average ($\overline{A}_t = 0.3A_t + 0.7\overline{A}_{t-1}$). If $|A_t - \overline{A}_t|$ exceeds a threshold, the cell is transient. We detect mover events only in stable cells, in which no transience has occurred for 2 s. Finally, a change in the cell before the incoming and after the outgoing motion is detected using a threshold on the difference in intensity gradient magnitudes, to filter out overall lighting changes over time.

Initial Hand Detection. When a mover event is detected, a region of size 90 × 90 around the detected cell is extracted as a presumed hand image and tracked for 2 s. From these regions in the *Movers* dataset (*SI Materials and Methods*), a total of 2,500 image patches containing movers are extracted and serve as positive examples to a standard classifier (28), using a nonparametric model variation (32). A total of 2,500 random image patches from surrounding regions serve as negative, nonclass examples. The resulting classifier is the initial hand detector, which is later extended automatically during the appearance–context iterations.

Appearance–Context Iterations. The initial hand detector learned from movers is automatically improved by combining two detection methods, one based on local appearance and the other on body context (28, 32). The two methods teach each other: hands detected by appearance provide a teaching signal for learning new context, and vice versa. All learning is completely unsupervised. The algorithm is applied in iterations, each iteration goes once over the full training data. This was used to compare performance attainable by internal teaching signals with full supervision. (In actual learning new data will be continuously supplied, which we tested, obtaining similar results.)

Each iteration performs the following. The highest scoring detections of the current detector are automatically chosen and tracked for up to 2 s (over which tracking is reliable). Image patches (90 × 90 pixels) around the tracked objects are used as presumed class examples. Patches from surrounding regions are used as nonclass examples. The existing appearance-based classifier (28) and a context-based classifier (32) are trained on the extracted patches. Briefly, the context-based classifier learns chains of features leading from the face to the hand through other body parts. The two new classifiers are used in conjunction by combining their detection scores, and the combined detector is used for training the next iteration.

Our experiments used a leave-one-out scheme, training on seven movies in the *Freely moving hands* dataset (*SI Materials and Methods*) and testing on the eighth unseen movie. This was repeated eight times, each time with a different test movie. Detection criteria increased over time: the first iteration was trained on the highest scoring 2% of detections by the initial movers-based detector. The second and third iterations used the top 10% and 20% of the previous detections, respectively. These percentages were determined once for all repetitions.

We compared our performance to fully supervised results. The supervised detector used the same algorithms for appearance and context but was trained on manually labeled hands from the entire training movies, using a similar leave-one-out scheme.

Gaze Detection. Gaze learning is done automatically, using mover events. Each detected mover event serves as a training example that assigns a presumed gaze direction to a face image. The face is detected at the onset of the event using an available face detector (48, 49) and represented using the HOG (Histogram-of-Gradients) descriptor (50). (If a face is not detected, the corresponding event is not used.) The gaze direction is taken as the direction from the center of the face produced by the detector, to the center of the cell that triggers the mover event. To estimate gaze direction in a novel image, the algorithm detects the face center, computes its HOG descriptor, and finds its K nearest neighbors in the training data, using L2 norm (K = 10 in all experiments). The predicted gaze direction is a weighted average of the gaze directions of the neighbors (weighted by similarity of the HOG descriptors).

Our experiments used a leave-one-out scheme, training on seven movies in the *Gaze* dataset (*SI Materials and Methods*) and testing on the eighth unseen movie. This is repeated eight times, each time with a different test movie, showing a different person.

For evaluation purposes we manually labeled events of object grasp and release, marking the target object at the frame of initial contact, and the center of the face. The direction from the face-center to the target is used as ground truth for measuring performance. The test set consists of patches of automatically detected faces in the labeled frames. For comparison, we evaluated human judgment of gaze direction: two observers were presented with the same test face patches as the algorithm and were asked to mark the gaze direction. Images on which either face detection failed or both human observers disagreed with the ground truth by more than 30° were removed from the test set. The final eight test sets contain 807 images in total (out of 887 initial contact images). We also compared the results with a supervised version of the algorithm using a similar leave-one-out procedure. It was trained on the manually labeled events of seven movies, using the same K nearest neighbors algorithm, and tested on the left-out movie.

ACKNOWLEDGMENTS. We thank M. Kloc for her early implementation of gaze extraction. The work was supported by European Research Council (ERC) Advanced Grant "Digital Baby" (to S.U.).

1. Woodward AL (1998) Infants selectively encode the goal object of an actor's reach. *Cognition* 69:1–34.
2. Gergely G, Bekkering H, Király I (2002) Rational imitation in preverbal infants. *Nature* 415:755–756.
3. Saxe R, Tenenbaum JB, Carey S (2005) Secret agents: Inferences about hidden causes by 10- and 12-month-old infants. *Psychol Sci* 16:995–1001.
4. Jordan MI (1998) *Learning in Graphical Models* (MIT Press, Cambridge, MA).
5. MacKay DJC (2003) *Information Theory, Inference, and Learning Algorithms* (Cambridge Univ Press, Cambridge, UK).
6. McClelland JL, et al. (2010) Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends Cogn Sci* 14:348–356.
7. Hinton GE (2007) Learning multiple layers of representation. *Trends Cogn Sci* 11: 428–434.
8. Geisler WS (2008) Visual perception and the statistical properties of natural scenes. *Annu Rev Psychol* 59:167–192.
9. Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND (2011) How to grow a mind: Statistics, structure, and abstraction. *Science* 331:1279–1285.
10. Fleischer F, Christensen A, Caggiano V, Thier P, Giese MA (2012) Neural theory for the perception of causal actions. *Psychol Res* 76:476–493.
11. 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.
12. Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. *Nat Neurosci* 5:682–687.
13. Walther DB, Koch C (2006) Modeling attention to salient proto-objects. *Neural Netw* 19:1395–1407.
14. Yoshida H, Smith LB (2008) What's in view for toddlers? Using a head camera to study visual experience. *Infancy* 13:229–248.
15. Aslin RN (2009) How infants view natural scenes gathered from a head-mounted camera. *Optom Vis Sci* 86:561–565.
16. Frank M, Vul E, Saxe R (2012) Measuring the development of social attention using free-viewing. *Infancy* 17:355–375.
17. Falck-Ytter T, Gredebäck G, von Hofsten C (2006) Infants predict other people's action goals. *Nat Neurosci* 9:878–879.
18. Pinker S (1997) *How the Mind Works* (W.W. Norton & Co., New York).
19. Spelke ES, Kinzler KD (2007) Core knowledge. *Dev Sci* 10:89–96.
20. Carey S (2009) *The Origin of Concepts* (Oxford Univ Press, New York).
21. Meltzoff AN, Kuhl PK, Movellan J, Sejnowski TJ (2009) Foundations for a new science of learning. *Science* 325:284–288.
22. Kremenitzer JP, Vaughan HG, Jr., Kurtzberg D, Dowling K (1979) Smooth-pursuit eye movements in the newborn infant. *Child Dev* 50:442–448.
23. Kaufmann F (1987) Aspects of motion perception in infancy. *Adv Psychol* 46:101–115.
24. Michotte A (1963) *The Perception of Causality* (Methuen, London).
25. Leslie AM (1984) Spatiotemporal continuity and the perception of causality in infants. *Perception* 13:287–305.
26. Premack D (1990) The infant's theory of self-propelled objects. *Cognition* 36:1–16.
27. Luo Y, Baillargeon R (2005) Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychol Sci* 16:601–608.
28. 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.
29. Slaughter V, Heron-Delaney M (2011) When do infants expect hands to be connected to a person? *J Exp Child Psychol* 108:220–227.

30. Slaughter V, Neary P (2011) Do young infants respond socially to human hands? *Infant Behav Dev* 34:374–377.
31. Amano S, Kezuka E, Yamamoto A (2004) Infant shifting attention from an adult's face to an adult's hand: a precursor of joint attention. *Infant Behav Dev* 27:64–80.
32. 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.
33. Scaife M, Bruner JS (1975) The capacity for joint visual attention in the infant. *Nature* 253:265–266.
34. Flom R, Lee K, Muir D (2007) *Gaze-Following: Its Development and Significance* (Lawrence Erlbaum Associates, Mahwah, NJ).
35. D'Entremont B, Hains SMJ, Muir DW (1997) A demonstration of gaze following in 3- to 6-month-olds. *Infant Behav Dev* 20:569–572.
36. Tomasello M (1999) *The Cultural Origins of Human Cognition* (Harvard Univ Press, Cambridge, MA).
37. Murphy-Chutorian E, Trivedi MM (2009) Head pose estimation in computer vision: A survey. *IEEE Trans Pattern Anal Mach Intell* 31:607–626.
38. Johnson M, Morton J (1991) *Biology and Cognitive Development: The Case of Face Recognition* (Blackwell, New York).
39. Sugita Y (2008) Face perception in monkeys reared with no exposure to faces. *Proc Natl Acad Sci USA* 105:394–398.
40. Bertenthal BI, Proffitt DR, Spetner NB, Thomas MA (1985) The development of infant sensitivity to biomechanical motions. *Child Dev* 56:531–543.
41. Sommerville JA, Woodward AL, Needham A (2005) Action experience alters 3-month-old infants' perception of others' actions. *Cognition* 96:B1–B11.
42. Piaget J (1952) *The Origins of Intelligence in Children* (W. W. Norton & Co., New York).
43. Rizzolatti G, Sinigaglia C (2008) *Mirrors in the Brain. How Our Minds Share Actions and Emotions* (Oxford Univ Press, New York).
44. Mori G, Ren X, Efros AA, Malik J (2004) Recovering human body configurations: Combining segmentation and recognition. *Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), pp 326–333, 10.1109/CVPR.2004.1315182.
45. 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.
46. Black MJ, Anandan P (1996) The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Comput Vis Image Underst* 63:75–104.
47. Bülthoff H, Little J, Poggio T (1989) A parallel algorithm for real-time computation of optical flow. *Nature* 337:549–555.
48. Bradski G (2000) The OpenCV Library. *Dr Dobbs J* 25:120, 122–125.
49. Conaire CO, O'Connor NE, Smeaton AF (2007) Detector adaptation by maximising agreement between independent data sources. *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), pp 1–6, 10.1109/CVPR.2007.383448.
50. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), pp 886–893, 10.1109/CVPR.2005.177.

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.

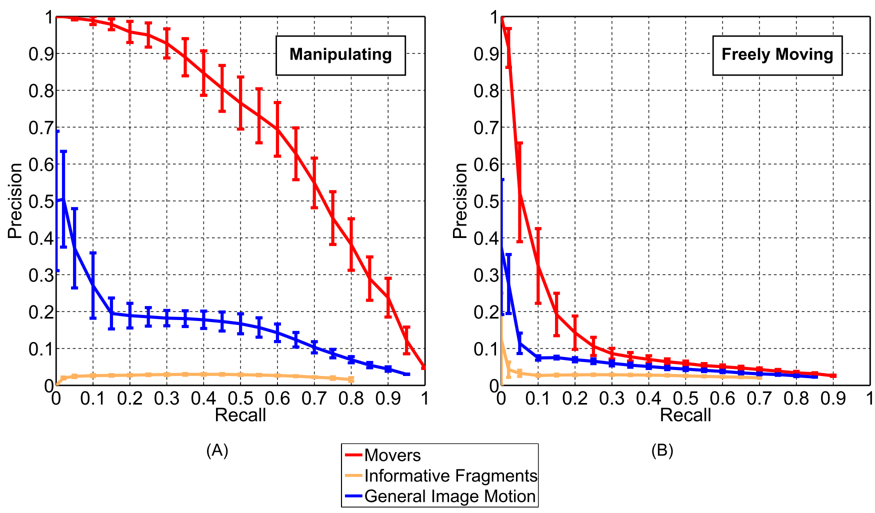
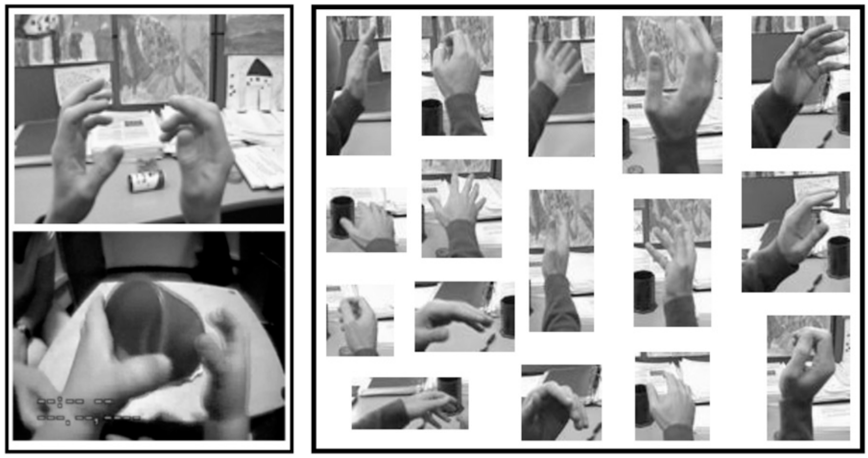


Fig. S1. Comparison of hand detectors trained by different cues. Precision-recall curves show mean and SE over all video sequences of (*A*) the *Manipulating hands* dataset, and (*B*) the *Freely moving hands* dataset. Red, training by mover events; blue, training by general motion; yellow, training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



(A) (B)

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.

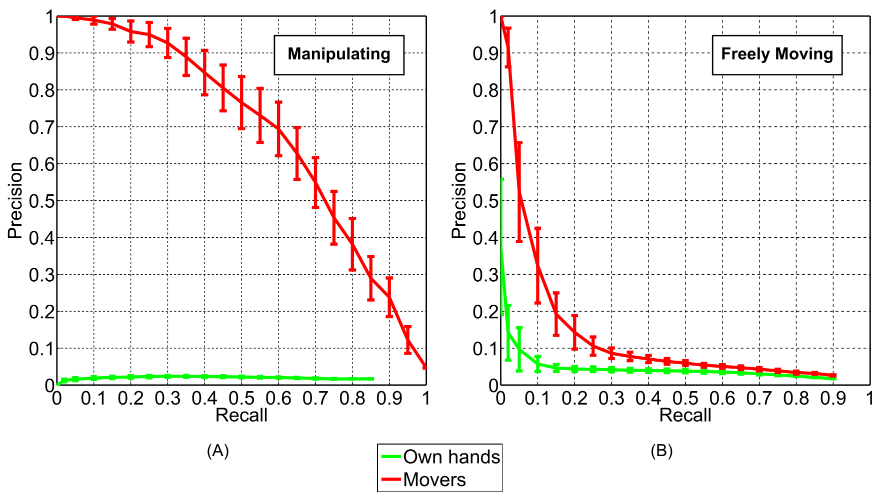


Fig. S3. Comparison of hand detectors trained by mover events and by own-hands. Precision-recall curves show mean and SE over all video sequences of (*A*) the *Manipulating hands* dataset, and (*B*) the *Freely moving hands* dataset. Red, training by mover events; green, training from own hands. Abscissa: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