Action as an innate bias for visual learning

Alan L. Yuille[a,b,1] and Heinrich H. Bülthoff[b,c,1]

[a]*Department of Statistics, University of California, Los Angeles, CA 90095;* [b]*Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, Korea; and* [c]*Department of Human Perception, Cognition and Action, Max Planck Institute for Biological Cybernetics, 72012 Tübingen, Germany*

Babies are faced at birth with a buzzing blooming confusion of visual stimuli (1). The set of all possible images is truly enormous (2), and simple calculations suggest that only a small fraction of all possible images have ever been seen over the entire history and prehistory of mankind. Moreover, the world consists of an estimated number of 30,000 objects (3), which occur in more than 1,000 different types of scenes. How can an infant start making sense of the visual world?

Detailed models of how infants learn to understand images and the balance between nature and nurture are currently lacking. Studies suggest that visual abilities develop in a stereotyped order (4). In particular, infants appear to be able to perceive motion and detect faces at an early stage of development. They can probably exploit the regularities that motion tends to be smooth in space and time, which also enables them to track image patches. Vision researchers have also demonstrated that many vertebrates and insects rely heavily on motion perception for surviving in this complex visual world, e.g., for camouflage breaking or figure ground separation (5, 6), and there are computational models that relate to neural circuitry (7).

Bootstrapping Visual Concepts

In PNAS, Ullman et al. (8) suggest a twist to this story by emphasizing the key role of motions that cause actions as innate biases that help bootstrap the learning of complex visual concepts. They address the phenomenon that infants are quick to learn models of hands and estimate the gaze direction of the owner of the hand. They point out that these are difficult tasks for which there are, as yet, no successful computational models (except those trained using supervision).

What makes a hand special for the infant? More precisely, how can the infant manage to isolate hands from the enormous amount of visual stimuli it perceives so that it can successfully learn to detect them? The motion of hands probably makes them interesting to the infant, but many other things move in images. Hands are associated with faces, to which infants are sensitive very early (as discussed later), but this is only indirect. The suggestion of Ullman et al. (8) is that a key property of hands is their ability to perform actions.

How does action help? Imagine a billiard table on which one ball is at rest and a second ball strikes it. The second ball acts on the first, transfers motion to it, and causes it to move. Similarly, in the infant's experience, a hand will often appear and act on a static object in the world, by moving a toy or offering a bottle. This distinguishes a hand from most moving objects in the infant's environment. Many objects can move, but those that also act by causing other objects to move are of particular importance. We emphasize that

Ullmann et al. suggest that visual learning by infants is partially driven by simple action events.

the infant probably has no concept of objects at this stage (with a few probable exceptions such as faces), and the theory of Ullman et al. (8) does not require it. Instead, they demonstrate that simple action events can be detected by local analysis of motion flow patterns. More precisely, by detecting events in which visual motion (e.g., the movement of a hand) flows into a previously static region of an image (e.g., a cup), motion flows out of the region (e.g., the hand carrying the cup), and then the region becomes static again (e.g., the part of the table on which the cup was resting).

Ullman et al. (8) implement a simple event detector and show that it will find hands but will also respond to other stimuli. Nevertheless, it is sufficient to isolate candidate regions of the image that can be exploited by tracking and modeling appearance and context (9). This helps bootstrap a stronger model, which is very effective at detecting hands. The context exploits the fact that fingers are linked to heads by a chain of body parts. Hence, the ability to detect heads can be used to help the detection of fingers by using a method developed by this research group (10). This model for hand detection, in turn, is used to learn a model of gaze perception, exploiting the fact that human gaze is frequently directed at hands.

Theory Predictions and Implications

The theory of Ullman et al. (8) raises some interesting questions. They have demonstrated the sufficiency of their computational model for learning hands and gaze directions. However, do infants use it? Their theory suggests experiments in which a new object (e.g., a billiard ball) that causes actions is introduced into the visual environment of an infant or a young monkey to see if these makes it easier to detect this new object. The theory would need to be extended to make concrete predictions in such situations. Issues like the exact timing of the action events, the presence or absence of context cues, and the amount of exposure required for learning would need to be explored.

To what extent are human abilities innate for these problems? Studies of monkeys suggest that they have some knowledge of faces before exposure. This has been shown by experiments in which animals have been raised without seeing faces until they are tested by behavioral experiments (11). Functional MRI studies (12) also suggest that face perception may be innate by contrast, for example, to the perception of text (13). This is not surprising, as monkeys and humans have been recognizing faces and interpreting facial expressions for more than hundreds of thousands of years, whereas humans have been reading text for a considerably shorter period. However, it is unclear that innate knowledge of faces extends to the ability to infer gaze direction directly from faces or to innate knowledge of hands (which have far greater variability than faces because of the articulation of fingers).

What other computational theories might be able to learn models of hands and gaze directions? As the authors state (8), most computational theories are based on the statistics of images and neglect concepts like action, which have a causal flavor, although recent work has extended statistics to include causality (14). The idea of using simple models to help bootstrap the learning of more

Author contributions: A.L.Y. and H.H.B. wrote the paper.

The authors declare no conflict of interest.

See companion article on page 18215.

[1]To whom correspondence may be addressed. E-mail: yuille@stat.ucla.edu or heinrich.buelthoff@tuebingen.mpg.de.

complex models has been applied with success to natural images (15) and, in particular, for learning objects by combining elementary parts together hierarchically to form complex objects (16). However, although compositional theories are conceptually attractive (17), there is, as yet, no clear evidence that humans use them. There have been surprisingly few computational models that exploit image sequences for learning object models in an unsupervised manner, with a few exceptions (18). This is despite the fact that behavioral studies show that the ability of adults to learn objects depends strongly on how views of the object appear in image sequences. In particular, the work by Wallis and Bülthoff (19) shows the importance of temporal associations for combining different views of an object into a single representation. These sequences of views can be obtained, for example, by moving the object by hand or by moving around it. Recently, computational modelers are starting to use motion sequences; e.g., Si et al. (20) have described a method for learning causality and actions from motion sequences. However, none of these methods, to our knowledge, have learned how to detect hands.

In any case, the use of action as a starting point for bootstrapping learning may initiate new directions of research. It also ties in, at the conceptual level, with other models of infant and child learning. The "theory-theory" (21) suggests that infants are like small scientists who learn by performing experiments on the world, seeking to understand its causal structure (e.g., dropping toys to explore gravity) and hence predict events. The ability to detect visual actions would seem fundamental to detecting the causal structures of events. A series of recent studies suggest that adults are often good at detecting the causal structure of visual scenes even from static images (22). Moreover, infants are good at estimating the causal structure of events of the billiard-ball launching type, and exhibit surprise when the normal causal rules are suspended by experimenters (23). Finally, of course, infants can control their own hands and are well situated to understand their causal properties.

In summary, Ullman et al. (8) suggest that visual learning by infants is partially driven by simple action events that help the infant pick out interesting parts of the enormous set of visual stimuli and use them to bootstrap up to complex models. From this perspective, human fondness for action movies, particularly those that contain action events, may reflect a highly successful learning mechanism rather than a mindless search for cheap thrills.

ACKNOWLEDGMENTS. The work was supported by the World Class University program through the National Research Foundation of Korea funded by Ministry of Education, Science and Technology Grant R31–10008.

1. James W (1890) *The Principles of Psychology with Introduction by George A. Miller* (Harvard Univ Press, Cambridge, MA).
2. Kersten D (1987) Predictability and redundancy of natural images. *J Opt Soc Am A* 4(12):2395–2400.
3. Biederman I (1987) Recognition-by-components: A theory of human image understanding. *Psychol Rev* 94(2): 115–147.
4. Kellman PJ, Arterberry M (1998) *The Cradle of Knowledge: Perceptual Development in Infancy* (MIT Press, Cambridge, MA).
5. Lettvin JY, Maturana HR, McCulloch WS, Pitts WH (1959) What the frog's eye tells the frog's brain. *Proceedings of the IRE* 47(11):1940–1951.
6. Bülthoff HH (1981) Figure-ground discrimination in the visual system of Drosophila melanogaster. *Biol Cybern* 41(2):139–145.
7. Reichardt W, Poggio T, Hausen K (1983) Figure-ground discrimination by relative movement in the visual system of the fly. Part II: Towards the neural circuitry. *Biol Cybern* 46(suppl):1–30.
8. Ullman S, Harari D, Dorfman N (2012) From simple innate biases to complex visual concepts. *Proc Natl Acad Sci USA* 109:18215–18220.
9. Oliva A, Torralba A (2007) The role of context in object recognition. *Trends Cogn Sci* 11(12):520–527.
10. 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, New York), pp 25–32.
11. Sugita Y (2008) Face perception in monkeys reared with no exposure to faces. *Proc Natl Acad Sci USA* 105(1):394–398.
12. Kanwisher N (2010) Functional specificity in the human brain: A window into the functional architecture of the mind. *Proc Natl Acad Sci USA* 107(25):11163–11170.
13. He S, et al. (2009) Transforming a left lateral fusiform region into VWFA through training in illiterate adults. *J Vis* 9(8):853.
14. Pearl J (2000) *Causality: Models, Reasoning, and Inference* (Cambridge Univ Press, New York).
15. Chen Y, Zhu LL, Yuille AL, Zhang H (2009) Unsupervised learning of probabilistic object models (POMs) for object classification, segmentation, and recognition using knowledge propagation. *IEEE Trans Pattern Anal Mach Intell* 31(10):1747–1761.
16. Zhu L, Lin C, Huang H, Chen Y, Yuille AL (2008) Unsupervised structure learning: hierarchical recursive composition, suspicious coincidence and competitive exclusion. *Proceedings of the European Conference on Computer Vision* (Springer, Berlin), Vol 5303, pp 759–773.
17. Geman S (2006) Invariance and selectivity in the ventral visual pathway. *J Physiol Paris* 100(4):212–224.
18. Wallraven C, Bülthoff HH (2001) Automatic acquisition of exemplar-based representations for recognition from image sequences. *Proceedings of the 2001 IEEE Conference on Computer Vision and Pattern Recognition Workshop on Models vs. Exemplars* (IEEE CS Press, Washington, DC).
19. Wallis GM, Bülthoff HH (2001) Effects of temporal association on recognition memory. *Proc Natl Acad Sci USA* 98(8):4800–4804.
20. Si Z, Pei M, Yao B, Zhu S-C (2011) Unsupervised learning of event and-or grammar and semantics from video. *Proceedings of IEEE International Conference on Computer Vision* (IEEE CS Press, Washington, DC), 41–48.
21. Gopnik A (2009) *The Philosophical Baby: What Children's Minds Tell Us about Truth, Love, and the Meaning of Life* (Farrar, Straus and Giroux, New York).
22. Hamrick J, Battaglia PW, Tenenbaum JB (2011). Internal physics models guide probabilistic judgments about object dynamics. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (Cognitive Science Society, Austin, TX), pp 1545–1550.
23. Leslie AM (1982) The perception of causality in infants. *Perception* 11(2):173–186.