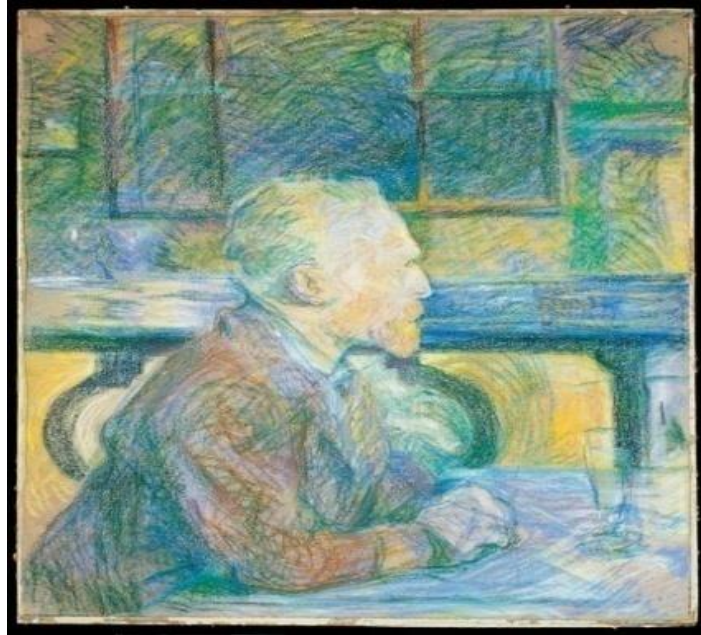


From simple innate biases to complex visual concepts



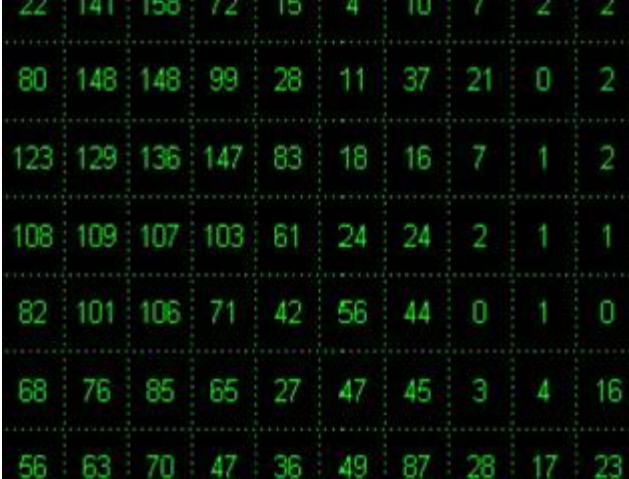
This image is in the public domain.



© Reuters. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

How it all starts

Image removed due to copyright restrictions.
Please see the video.



22	141	158	72	15	4	10	7	2	2
80	148	148	99	28	11	37	21	0	2
123	129	136	147	83	18	16	7	1	2
108	109	107	103	61	24	24	2	1	1
82	101	106	71	42	56	44	0	1	0
68	76	85	65	27	47	45	3	4	16
56	63	70	47	36	49	87	28	17	23

© Source Unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

- Start without world knowledge
- Watch many movies of the world
- Develop representations of various concepts



© Harry L Anthony. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

Hands



© ciifka at Flickr.com. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

Gaze

Difficult, appear early, important for subsequent learning of agents, goals, interactions,

Hands and body parts are important



© [Somesai](#) via Flickr.com. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

Action recognition
Gesture and communication
Agents interactions

Hands are difficult



© Source Unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

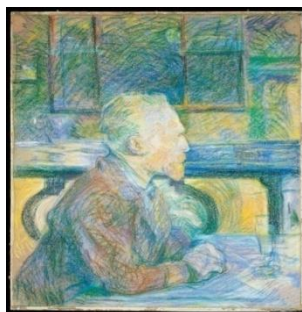
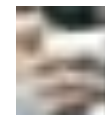
Multiple appearances



© Joe Amaro. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.



© Source Unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.



This image is in the public domain.

Van Gogh



© Ernst Kerchner. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

Kirchner

Small and inconspicuous

Difficult to extract in unsupervised schemes



© Source Unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

Informative fragments from people /
no-people

‘The problem of recovering human body configurations in a general setting is arguably the most difficult recognition problem in computer vision’

Figure removed due to copyright restrictions.
Please see the video.

Unsupervised Deep
Learning

Mori, Malik, CVPR 2004

Unsupervised learning does not discover hands

Building High-level Features Using Large Scale Unsupervised Learning
Ng et al Stanford and Google ICML 2012

Figure removed due to copyright restrictions. Please see the video.
Source: Le, Quoc V. "Building high-level features using large scale unsupervised learning." In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 8595-8598. IEEE, 2013.

1B connections, 10M YouTube images, 1000 machines, 16,000 cores, 3 days

Some statistically significant structures emerge with large data

In humans: Selectivity to hands appear early in infancy

Using a Head Camera to Study Visual Experience.



© Wiley. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.
Source: Yoshida, Hanako, and Linda B. Smith. "What's in view for toddlers? Using a head camera to study visual experience." *Infancy* 13, no. 3 (2008): 229-248.

‘Overall...hand were in view and dynamically acting on an object in over 80% of the frames’.

Yoshida & Smith 2008

What makes hands learnable by humans?

Motion, Hand as 'mover' (7-months old)

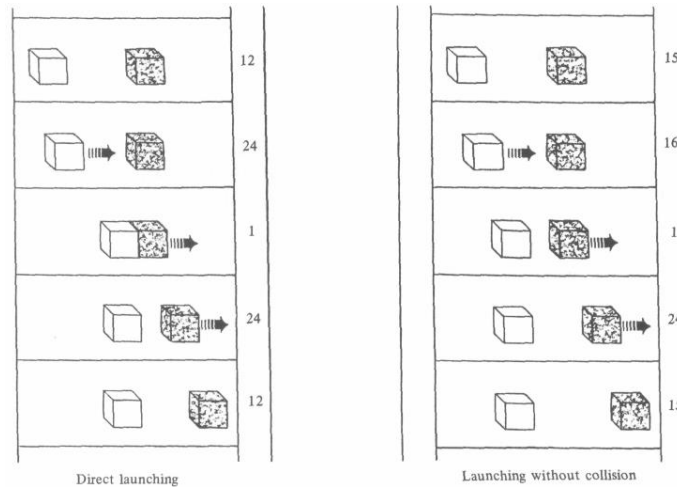


© fotosearch. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

See: Saxe, Carey The perception of causality in infancy. *Acta Psychologica* 2006

Early sensitivity to special motion types

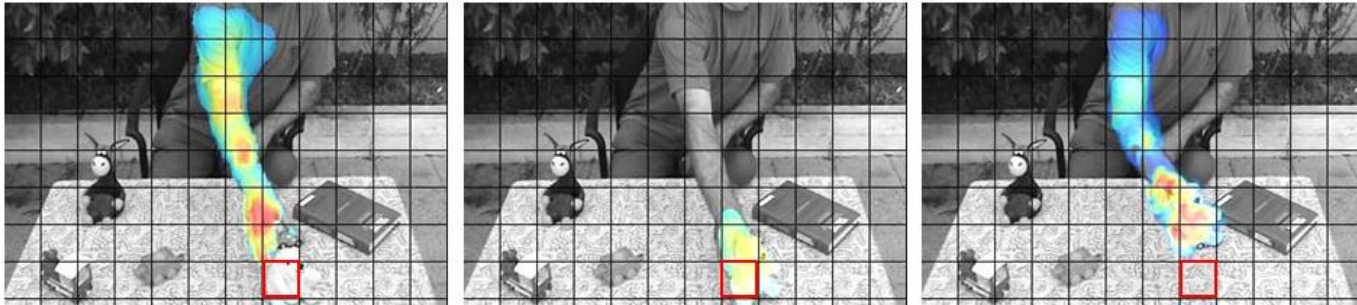
- High sensitivity to motion in general
(detecting motion, motion segmentation, tracking)
- Specific sub-classes of motion: self-motion, passive, and 'mover'



© Source Unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

A specific motion even is highly indicative of hands

Detecting 'Mover' Events



Courtesy of National Academy of Sciences, U. S. A. Used with permission.
Source: Ullman, Shimon, Daniel Harari, and Nimrod Dorfman. "From simple innate biases to complex visual concepts." *Proceedings of the National Academy of Sciences* 109, no. 44 (2012): 18215-18220. Copyright © 2012 National Academy of Sciences, U.S.A.

A moving image region causing a stationary region to move or change after contact.

Simple and primitive, prior to objects or figure-ground segmentation

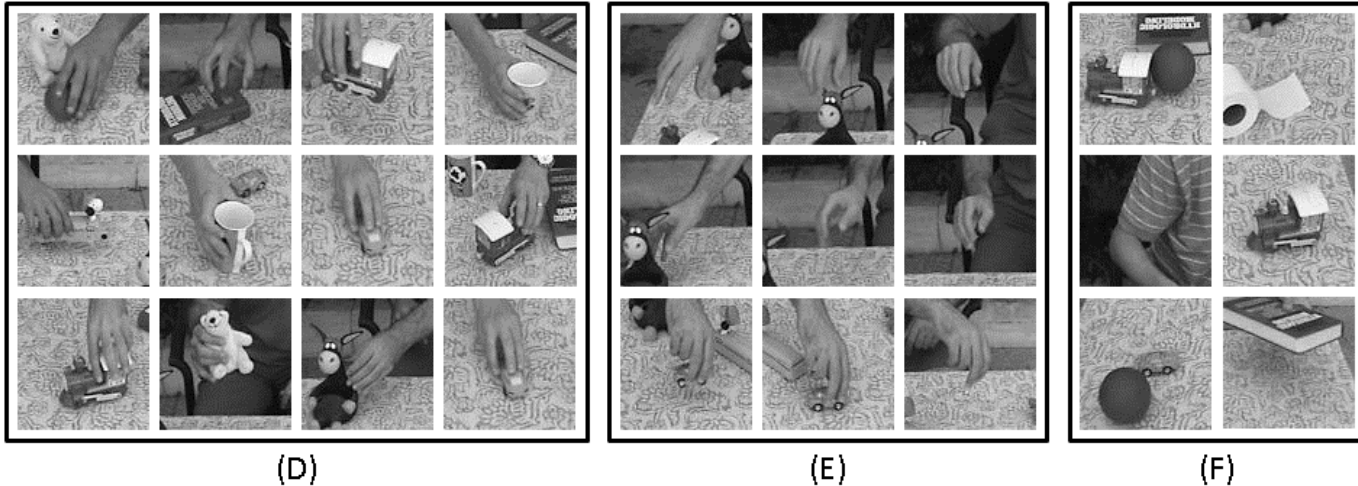
Movers detection



'Mover' as an innate teaching signal for hand

Motion alone is insufficient

'Mover' events extracted from videos

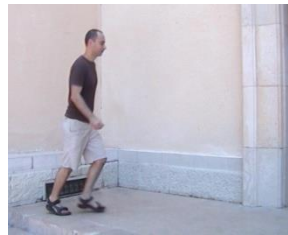


Courtesy of National Academy of Sciences, U. S. A. Used with permission.
Source: Ullman, Shimon, Daniel Harari, and Nimrod Dorfman. "From simple innate biases to complex visual concepts." *Proceedings of the National Academy of Sciences* 109, no. 44 (2012): 18215-18220. Copyright © 2012 National Academy of Sciences, U.S.A.

High fraction of Hand images
(90% recall 65% precision)

Internal supervision by movers and by tracking

Training Videos



Movies of scenes, people moving, manipulating objects, moving hands.

‘Mover’ events are detected in all movies and used for training

Hand detection in still images



© Proceedings of the National Academy of Sciences. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

Source: Ullman, Shimon, Daniel Harari, and Nimrod Dorfman. "From simple innate biases to complex visual concepts." Proceedings of the National Academy of Sciences 109, no. 44 (2012): 18215-18220.

Detection mainly of hands in object manipulation scenes

Continued learning

- Two detection algorithms:

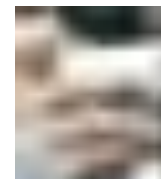
- Hands by their appearance



- Hands by the body context

Figure removed due to copyright restrictions. Please see the video.
Source: Karlinsky, Leonid, Michael Dinerstein, Daniel Harari, and Shimon Ullman. "The chains model for detecting parts by their context." In *Computer Vision and Pattern Recognition (CVPR)*, 2010 IEEE Conference on, pp. 25-32. IEEE, 2010.

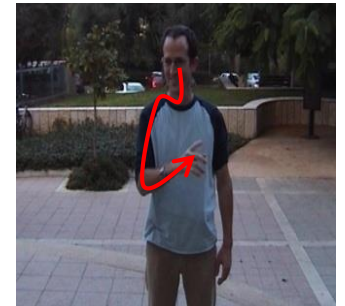
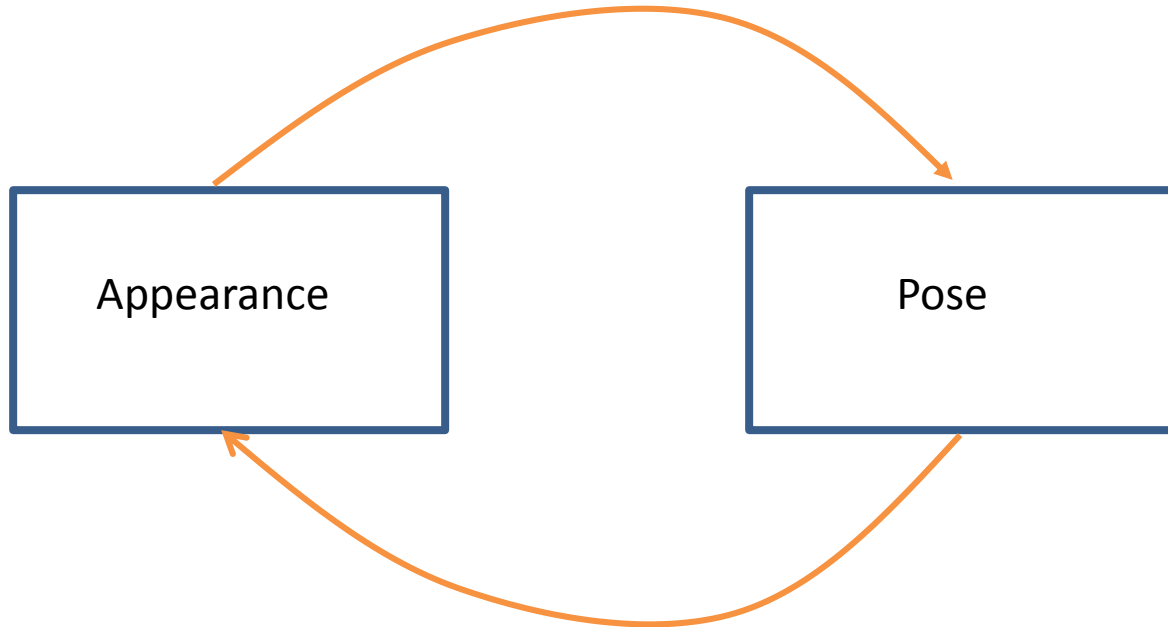
Hand by Surrounding Context



Face → Shoulder → Upper-arm → Lower-arm → Hand

Amano, Kezuka, Yamamoto 2004
Slaughter Heron-Delaney 2010
Slaughter, Neary 2011

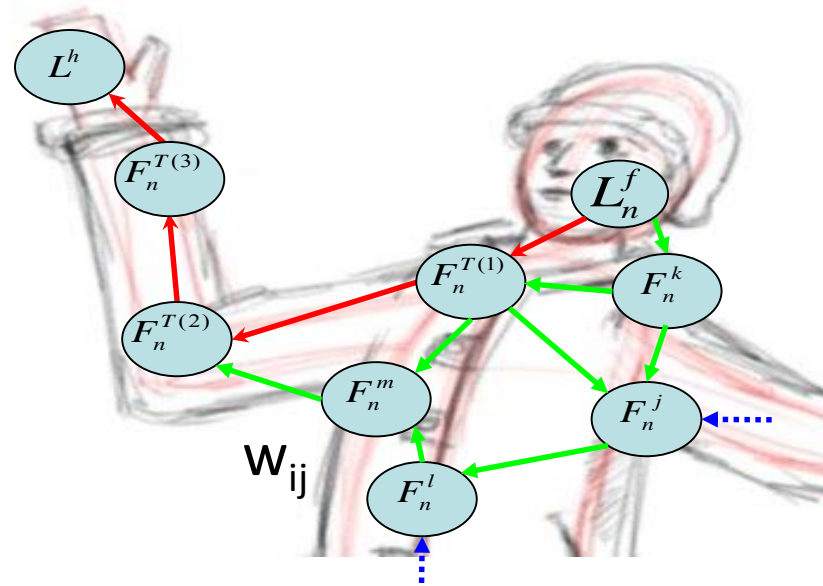
Co-training



Two supervised classifiers
Internal co-supervision

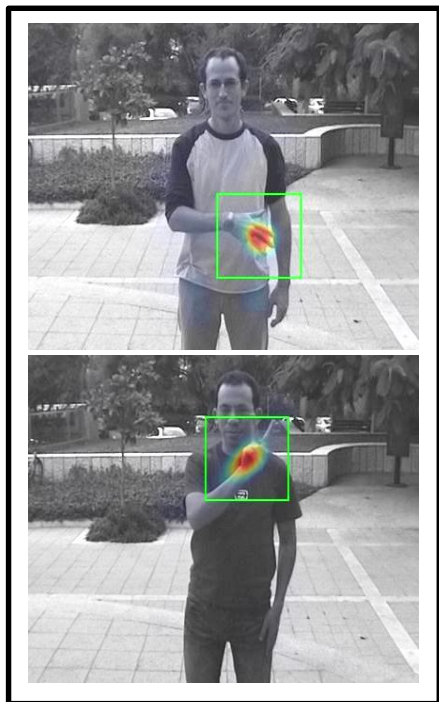
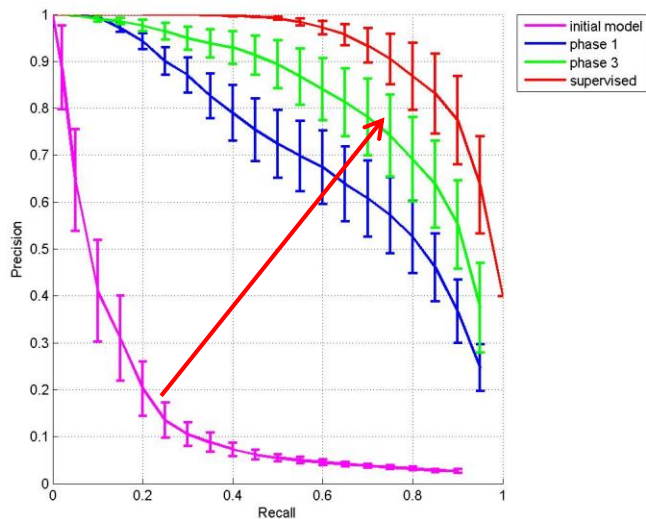
The chains computation:

Chains model



© The Weizmann Institute. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

(a)



(c)



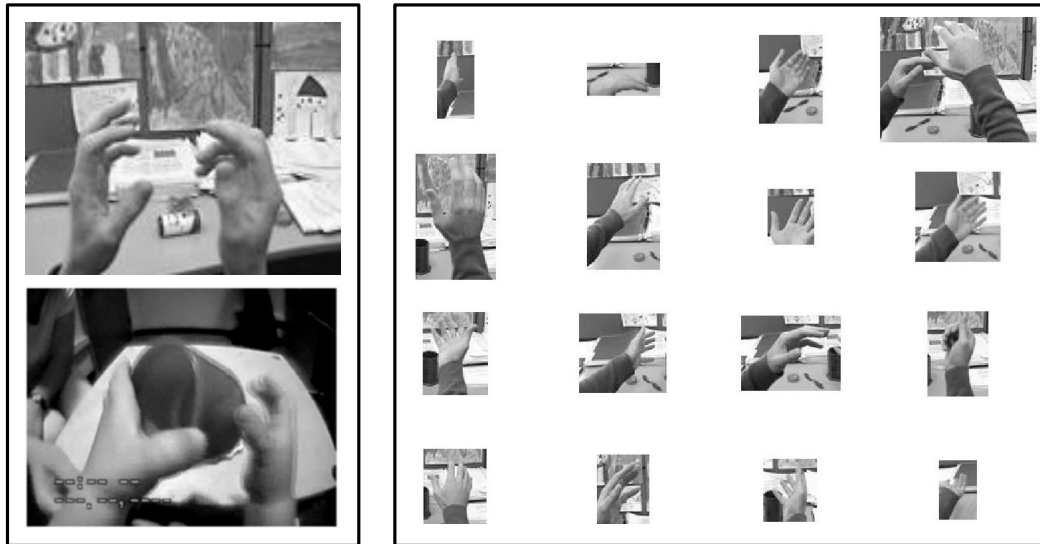
(d) Appearance



(e) Context

Courtesy of National Academy of Sciences, U. S. A. Used with permission.
Source: Ullman, Shimon, Daniel Harari, and Nimrod Dorfman. "From simple innate biases to complex visual concepts." Proceedings of the National Academy of Sciences 109, no. 44 (2012): 18215-18220. Copyright © 2012 National Academy of Sciences, U.S.A.

Own Hands



(A)

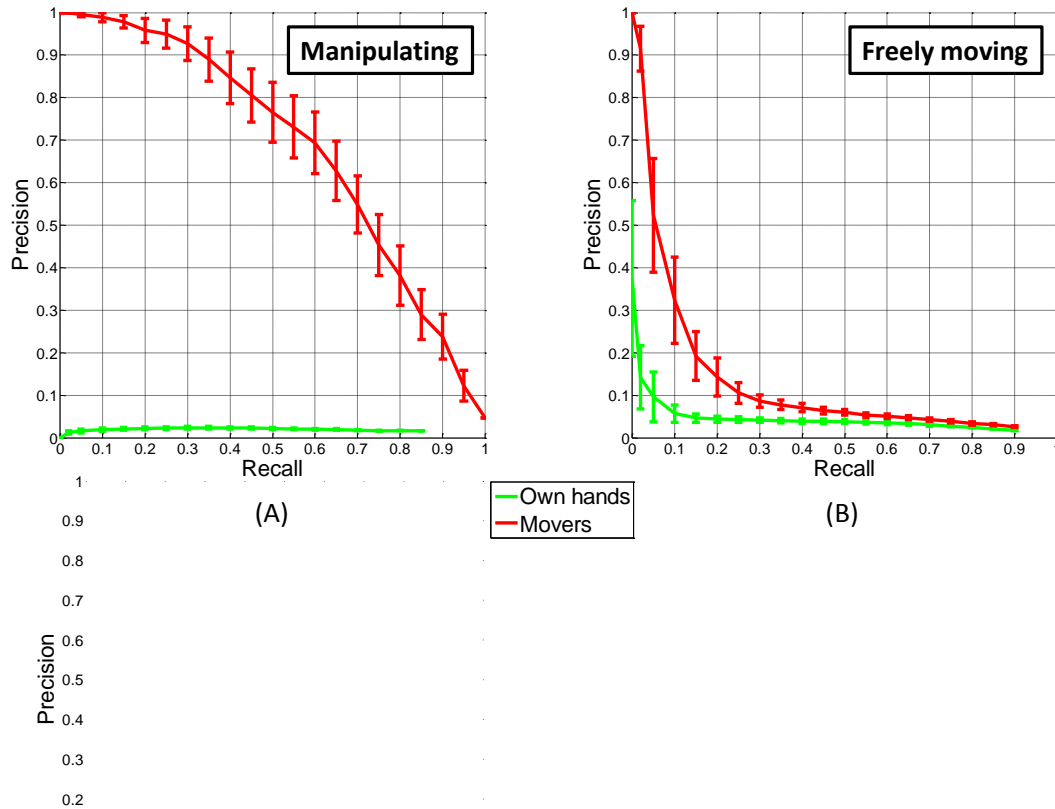
(B)

© Wiley. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.
Source: Yoshida, Hanako, and Linda B. Smith. "What's in view for toddlers? Using a head camera to study visual experience." *Infancy* 13, no. 3 (2008): 229-248.

Yoshida & Smith

A learned class, not the basis of hands in general
Caregiver's hands

Own Hands



Gaze



© [ciifka](#) at Flickr.com. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

Infants follow the gaze of others

Starting at 3-6 months and continues to develop

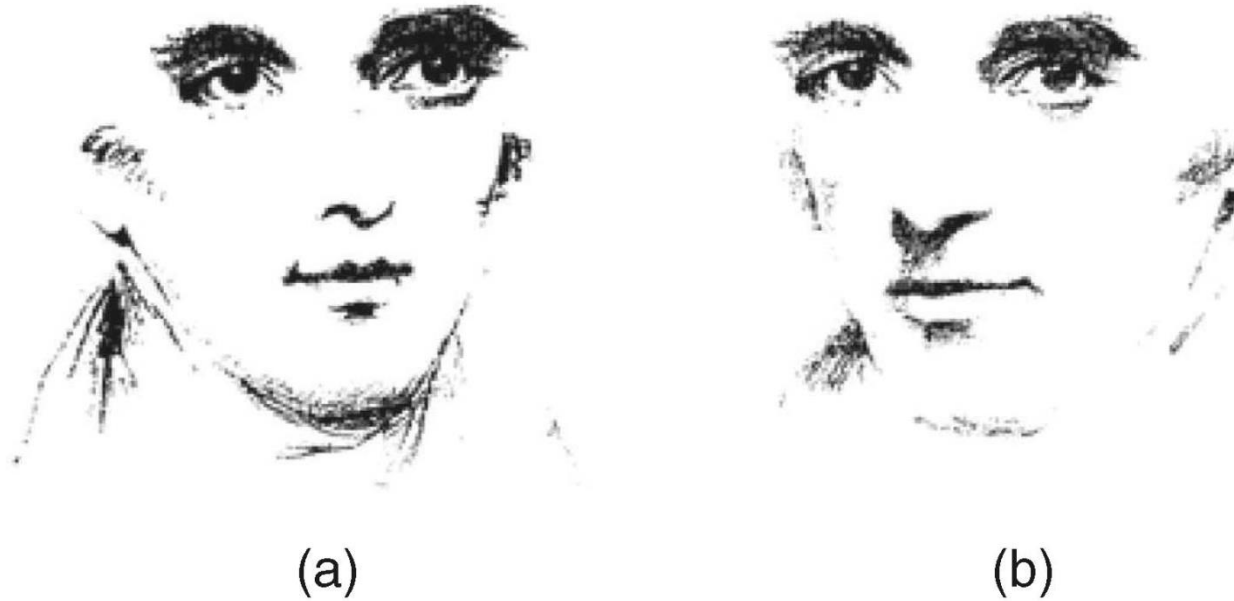
Head orientation first, eye cues later

Important in the development of communication and language

Modeling mainly head direction



Wollaston 1824



This image is in the public domain.

W.H. Wollaston, "On the Apparent Direction of Eyes in a Portrait," *Philosophical Trans. Royal Soc. of London*, 1824.

Gaze cues are subtle and inconspicuous



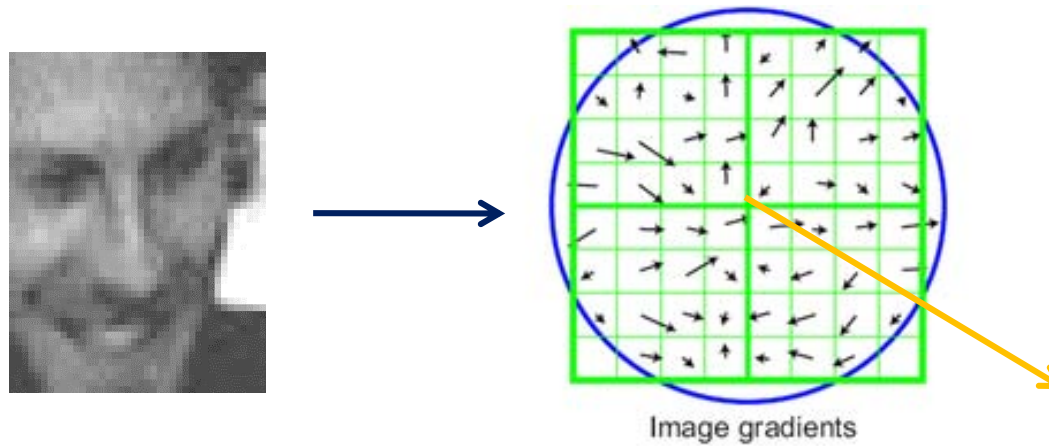
Mover supplies the teaching signal



Using hand 'mover' events to learn gaze direction



HoG description

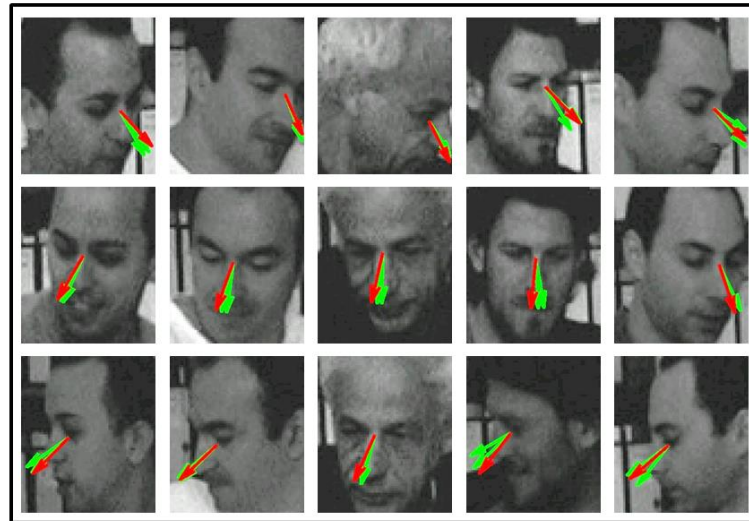


Gaze extraction 2D



(D)

Training



(E)

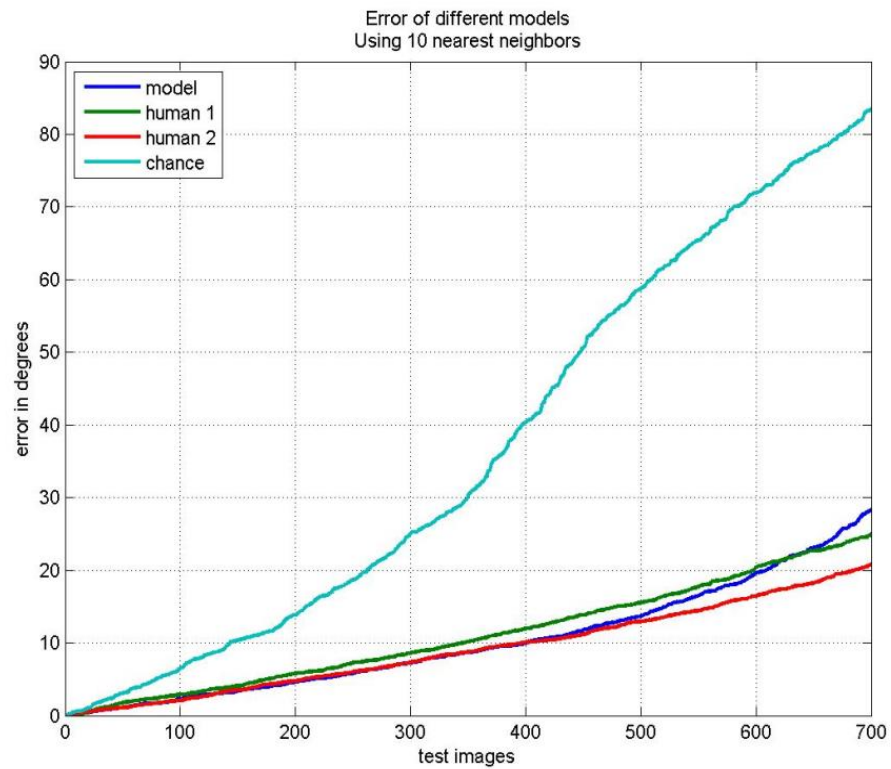
Testing

Model 

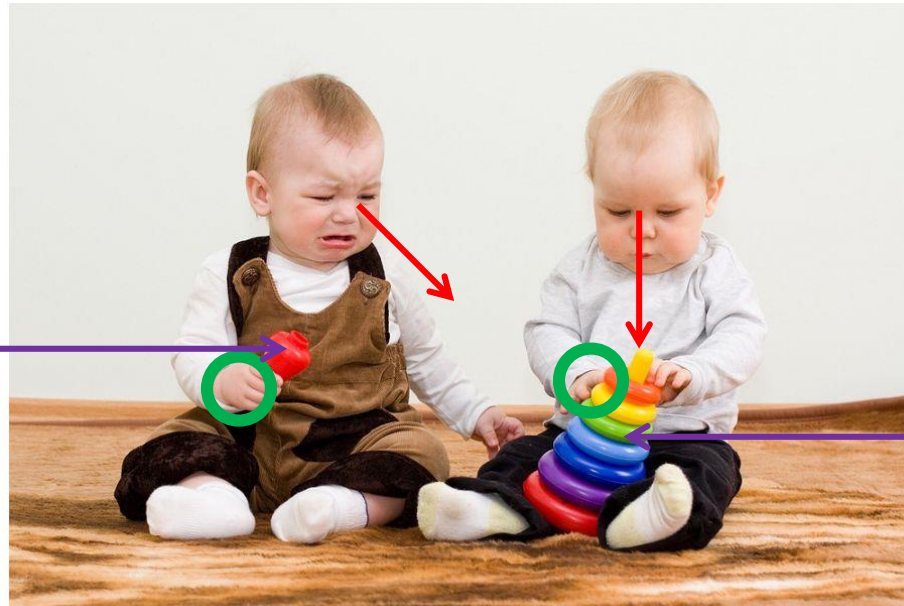
Humans 

Gaze results, 700 test images

8 people, leave-one-out



Emerging Interpretation

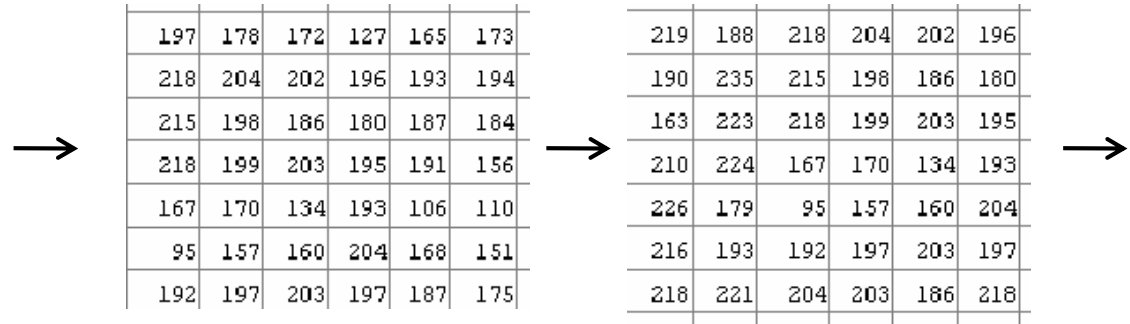


© Shutterstock. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

Both agents are manipulating objects;
The one on the left is interested in the other's object

'Digital Baby'

Innate capacities
 Mover
 Tracking
 Mover-to-gaze
 Co-training



184	113	118	105	117	82	:
151	95	122	131	87	100	:
160	156	159	197	178	172	:
136	219	188	218	204	202	:
184	190	235	215	198	186	• • •
175	163	223	218	199	203	:
221	210	224	167	170	134	:

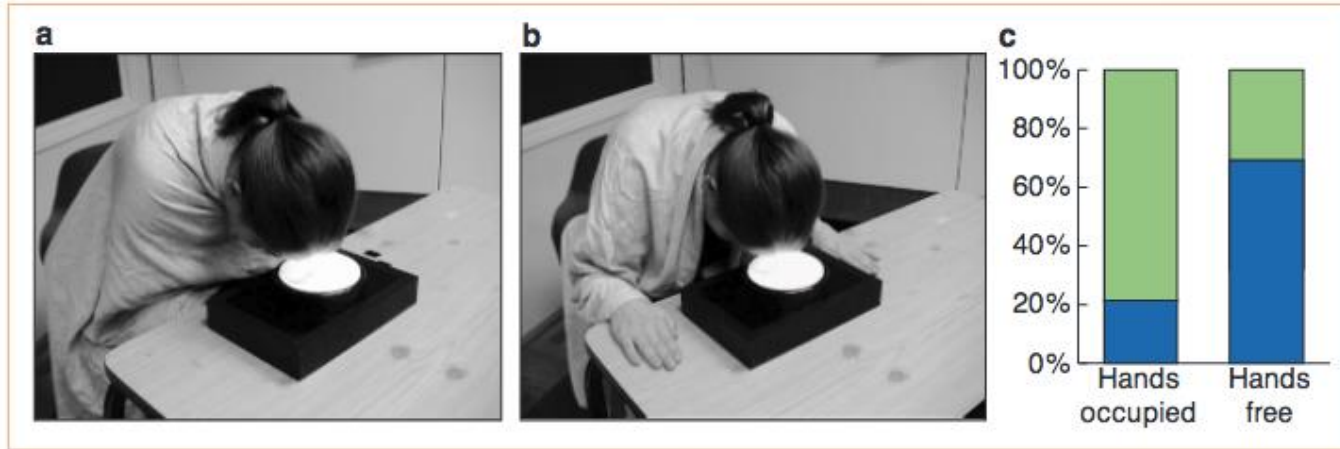
Figure removed due to copyright restrictions. Please see the video.
 Source: Karlinsky, Leonid, Michael Dinerstein, Daniel Harari, and Shimon Ullman. "The chains model for detecting parts by their context." In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 25-32. IEEE, 2010.



Concepts
 Hand – appearance
 Hand – context
 Gaze
 Nouns, verbs

© Source Unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

Rational imitation in preverbal infants



Reprinted by permission from Macmillan Publishers Ltd: Nature.
Source: Gergely, György, Harold Bekkering, and Ildikó Király. "Developmental psychology: Rational imitation in preverbal infants." *Nature* 415, no. 6873 (2002): 755. © 2002.

Gyorgy Gergely, Harold Bekkering, Ildiko Kiraly, *Nature* 415, 2002

Learning and innate structures

- Complex concept neither learned on its own nor innate.
- Domain-specific innate structures
- Not full solutions, but proto-concepts and strategies
- Not hands, but movers etc.
- Guide the system to develop meaningful representations
- Provide internal supervision
- ‘Learning trajectories’: mover – hand – gaze – reference
- Can extract meaningful concepts event when they are non-salient in the input
- From cognition to AI: incorporate similar structures in computational systems

MIT OpenCourseWare
<https://ocw.mit.edu>

Resource: Brains, Minds and Machines Summer Course
Tomaso Poggio and Gabriel Kreiman

The following may not correspond to a particular course on MIT OpenCourseWare, but has been provided by the author as an individual learning resource.

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.