UNIVERSITY OF MARYLAND 18 56

NATIONAL INSTITUTES OF HEALTH

poeticon++

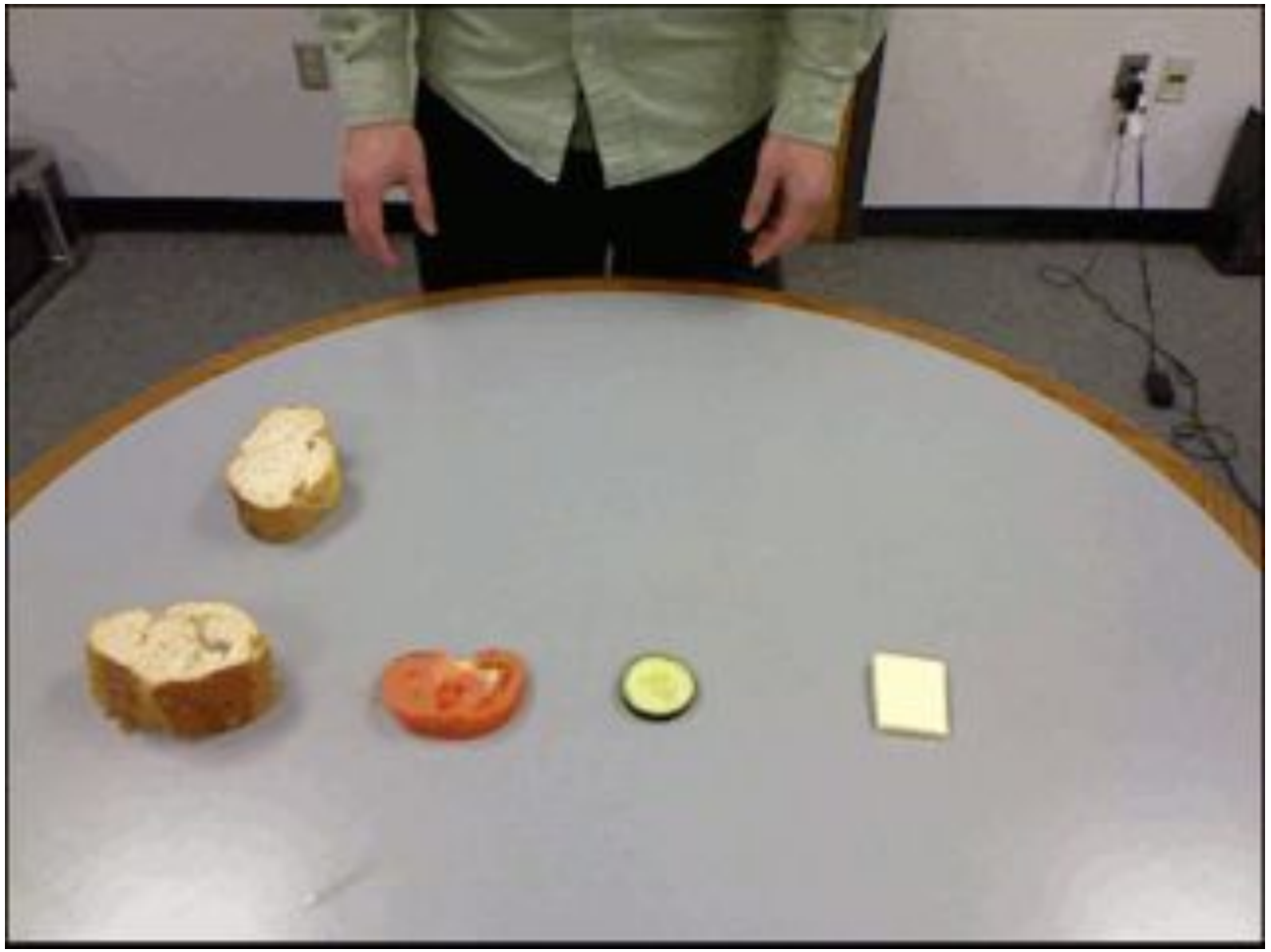
NSF

QUALCOMM®

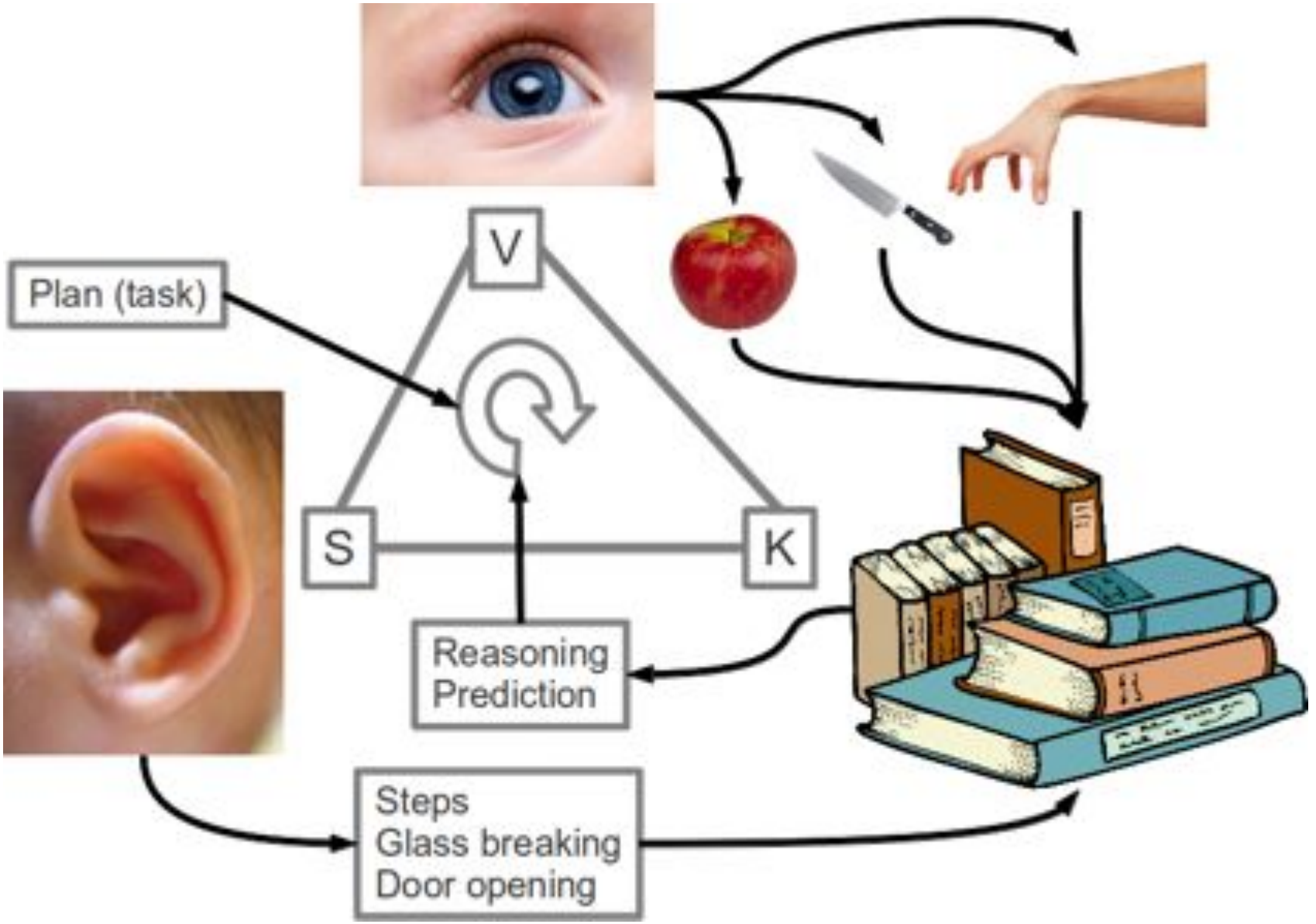
Degrees of freedom problem

Perception-Action integration

Action Sequencing



Learning


V
Plan (task)
S
K
Reasoning
Prediction
Steps
Glass breaking
Door opening

Actions contain

Objects, tools, hands, movements

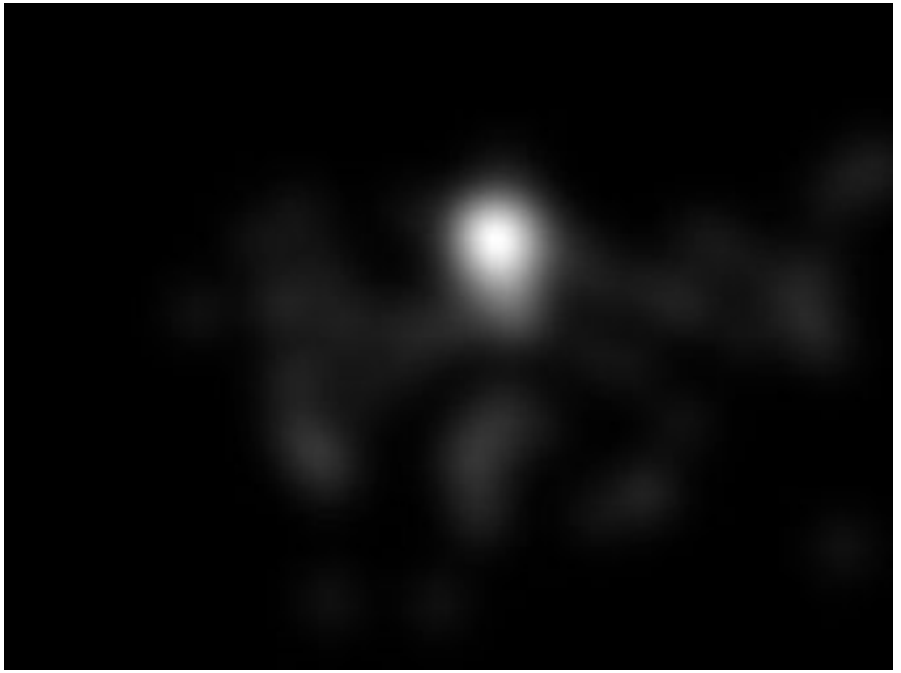
Plan of the presentation

1. Attention, finding/recognizing objects using contours.

2. Recognizing goals (action consequences).

3. Parsing action sequences using the manipulation grammar.

The Bottom up Attention Mechanism

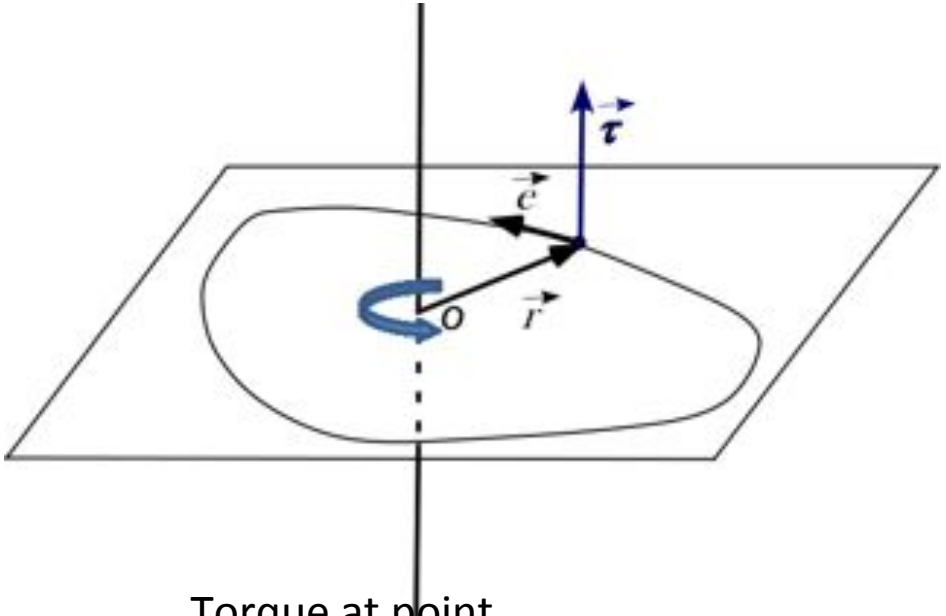


A Motivating Example

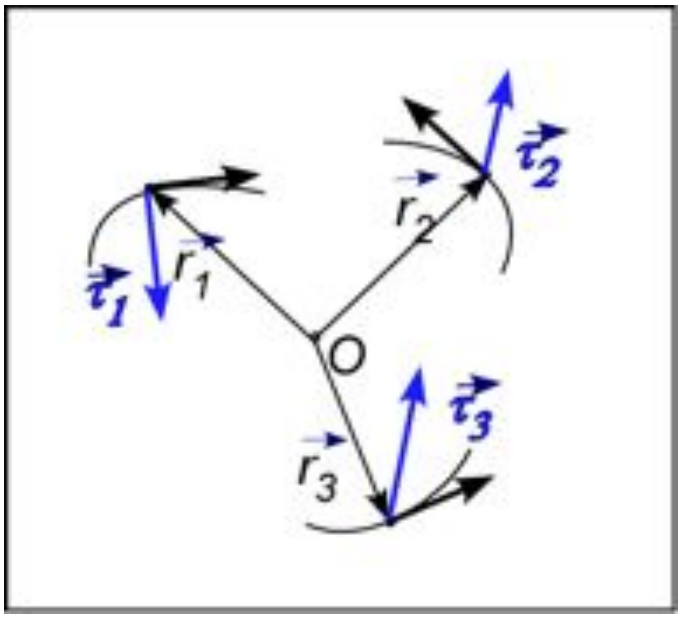
Where are my scissors?

The torque

- encodes enclosed, connected regions



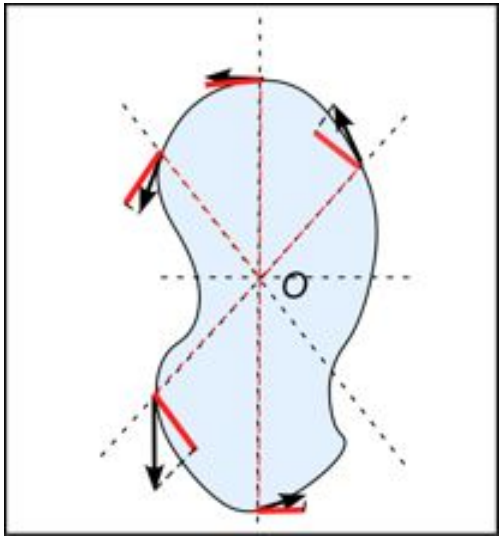
Torque at point



Torque in a patch

Value of torque : r x c

Intuition for the torque

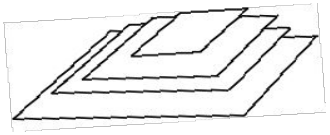


Torque (Patch) : Sum of radial components weighted by distance from center
encodes the enclosed **area**

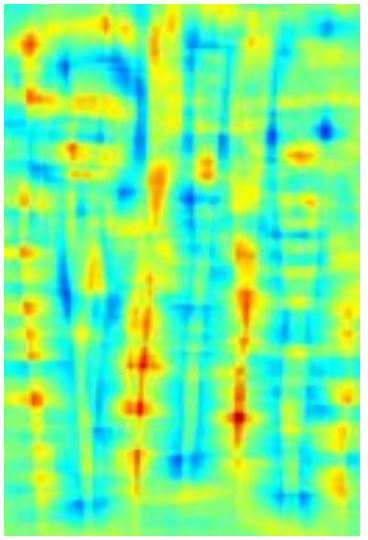
Properties:

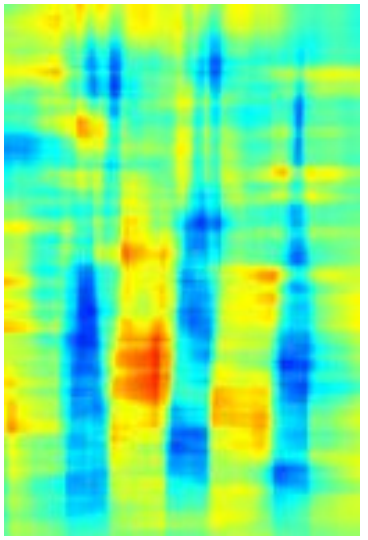
- Large for enclosed regions Small for random texture
- Large for contours centerd around origin
- Edges are oriented : by gradient (single images)
  depth cues (borderownership) (motion and stereo)

Torque map



Torque at multiple Window sizes



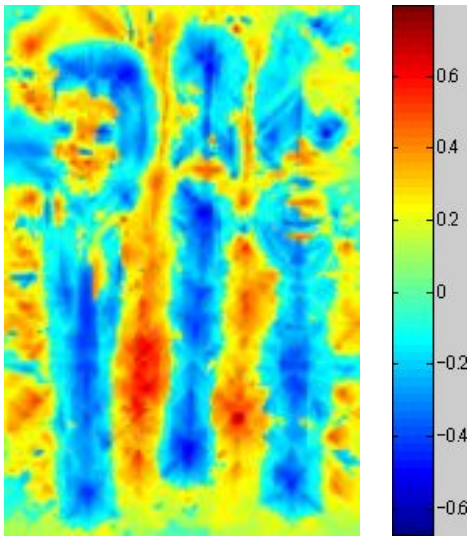


size = 5 x 5

21 x 21

45 x 45

Combined Torque map



Extrema over scale

Canny edges

Use torque map to strengthen edges

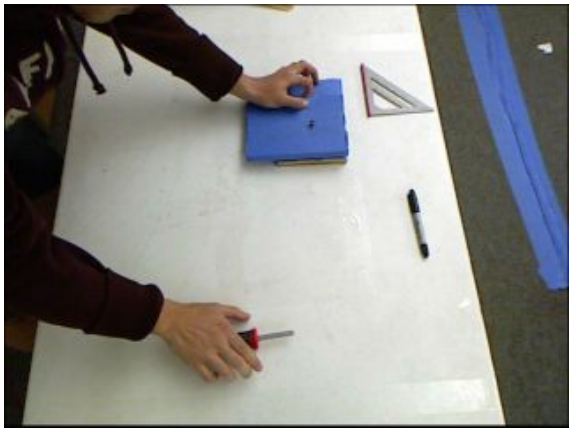
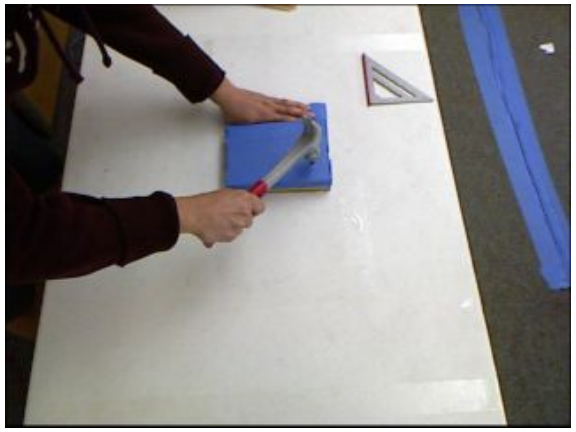
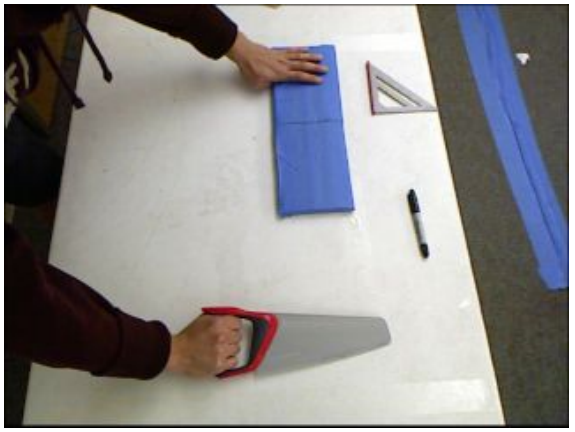
Object Recognition in clutter/occlusion

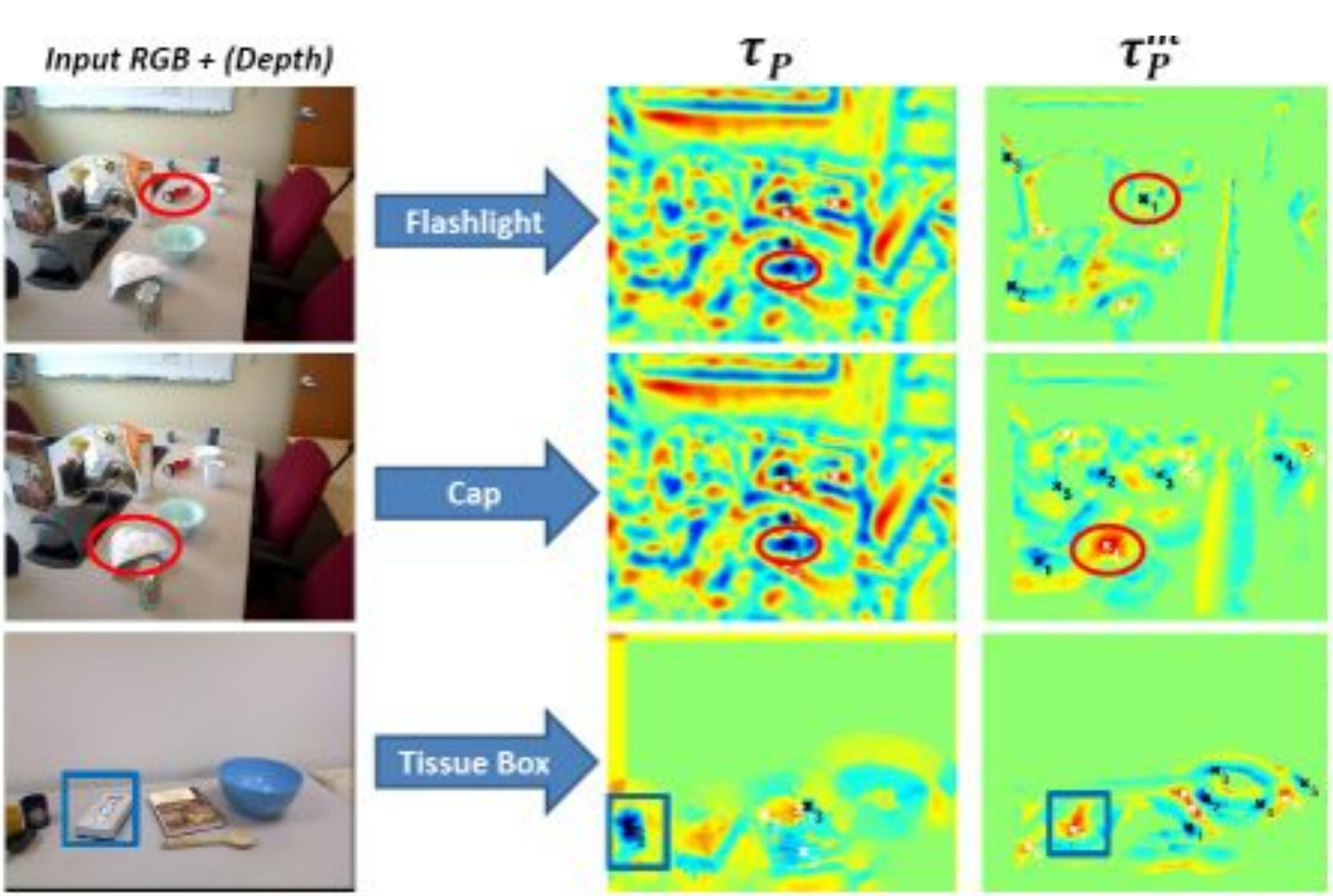
Typical scenes from manipulation/search tasks:

1) Clutter with partial occlusions and viewpoint changes

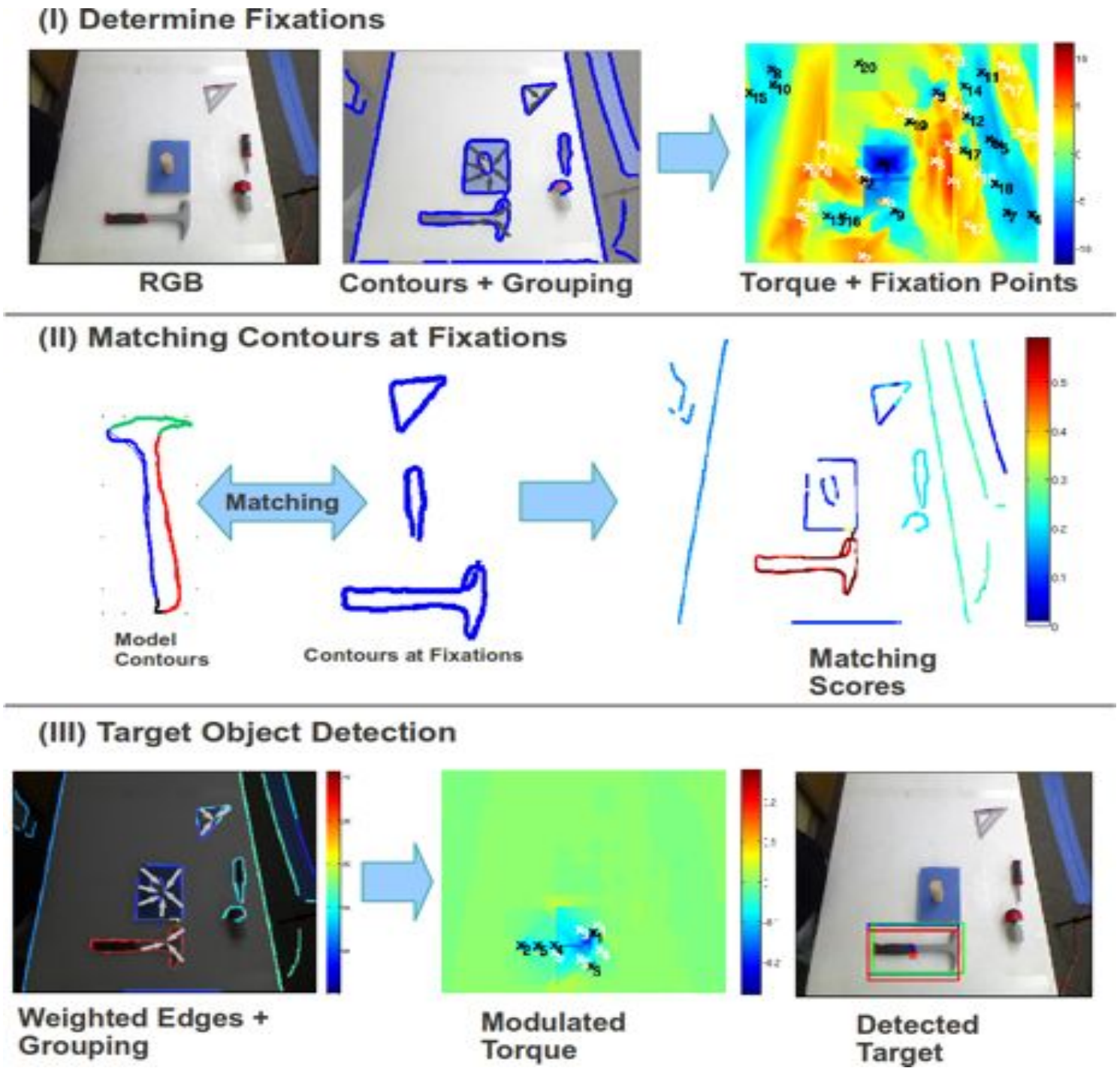


2) Occlusions due to hands, objects in various orientations

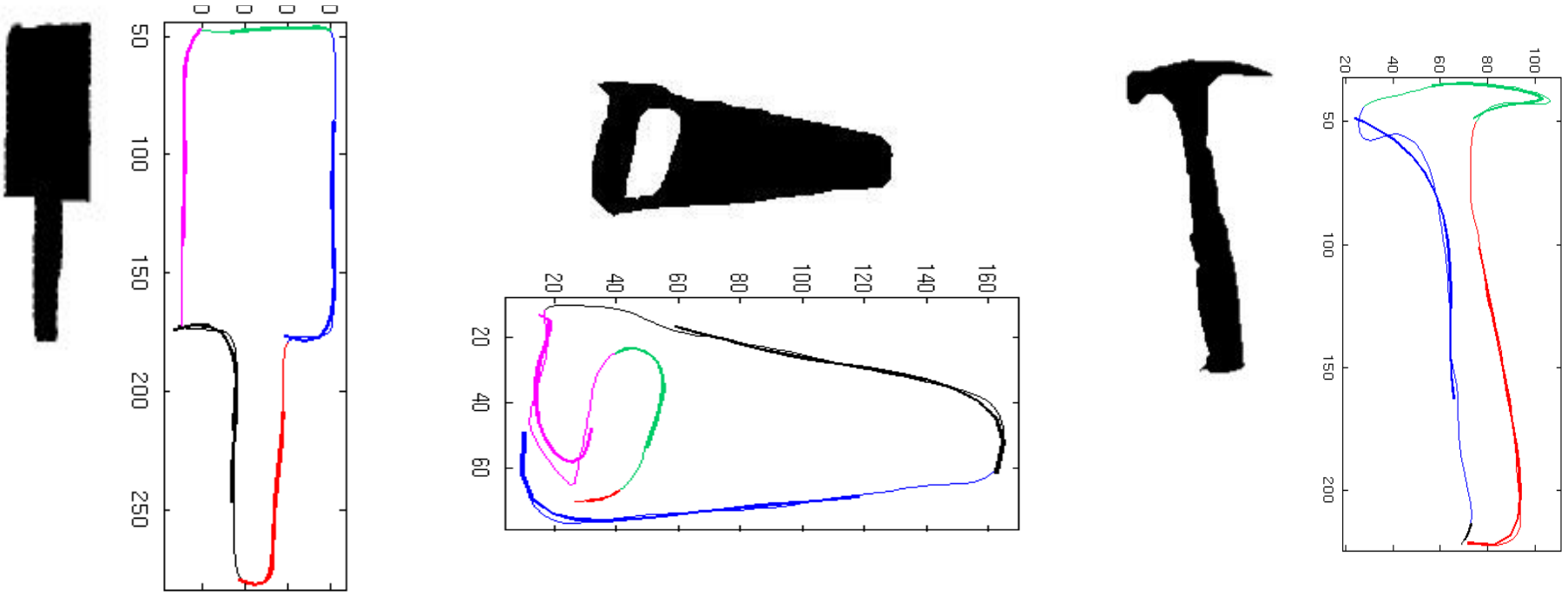





Input RGB + (Depth)
τ_P
τ_P^m
Flashlight
Cap
Tissue Box

Solution (3 steps)



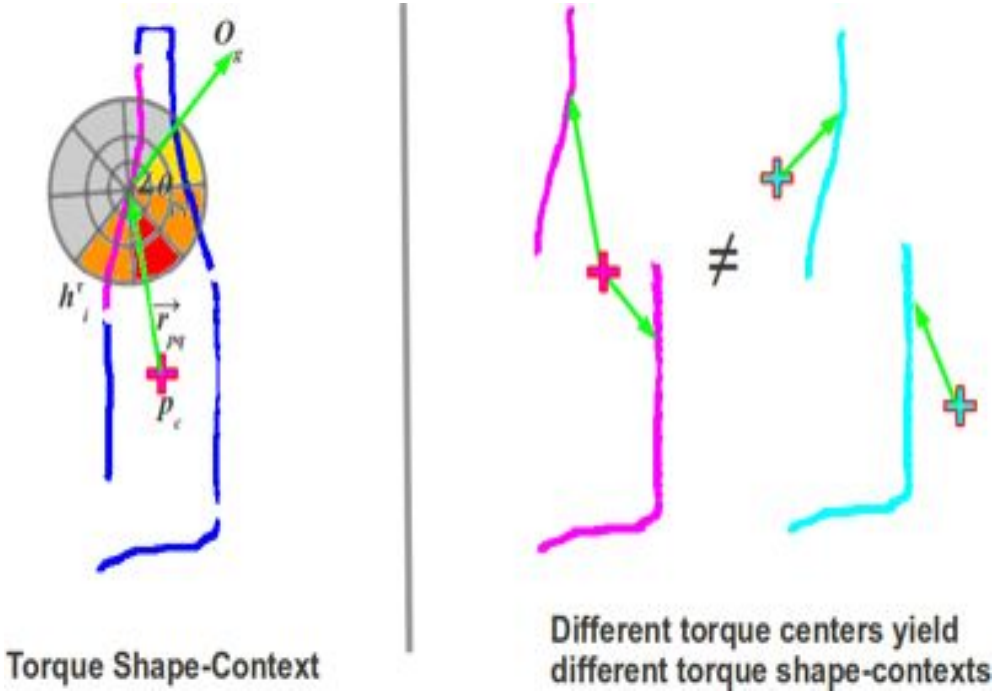
- **Learn** salient contours from annotated examples
- **Match** partial contour fragments to models
- **Reweigh** torque based on matching scores
- Learned contours examples:

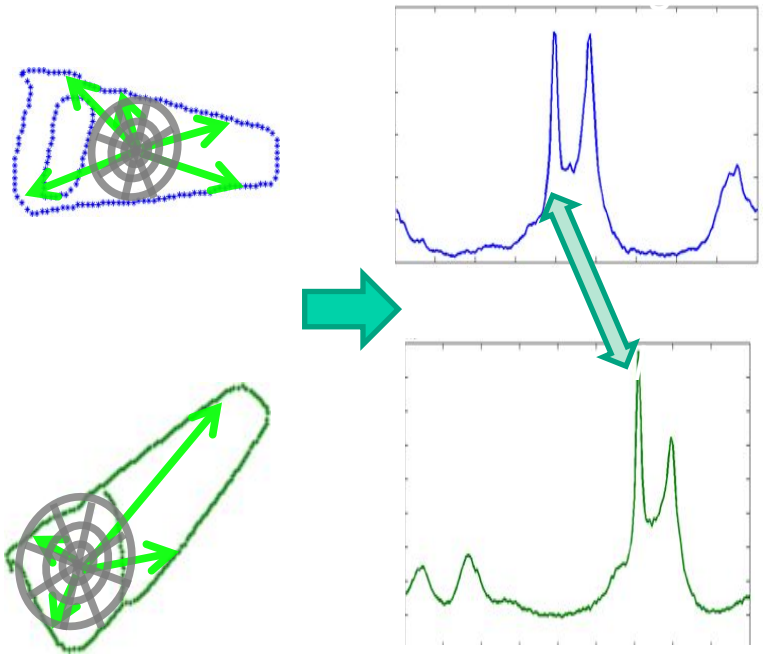


Novel Edge Fragment Descriptor

Robust fragment matching via torque + shape context

Rotational invariance from FFT of descriptor



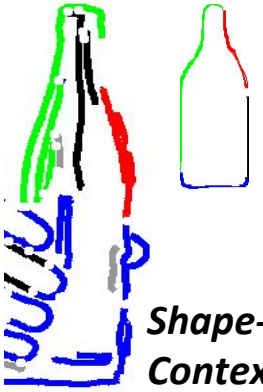


Modulate Torque for Object Similarity

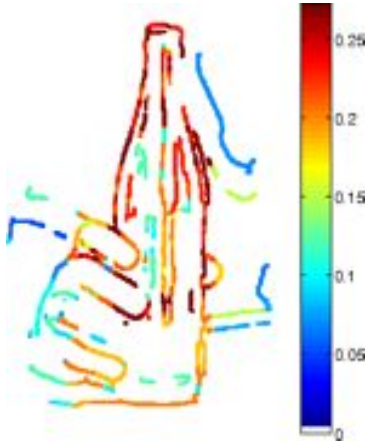
Shape Context Matching with Torque Centers

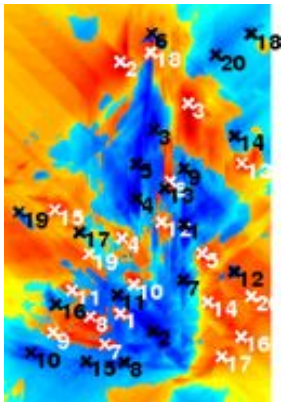
Matching

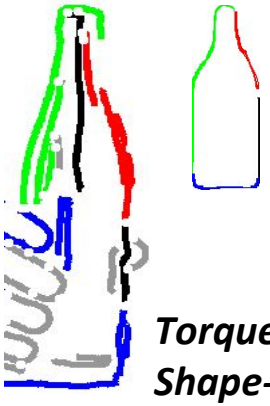
Edge Weights



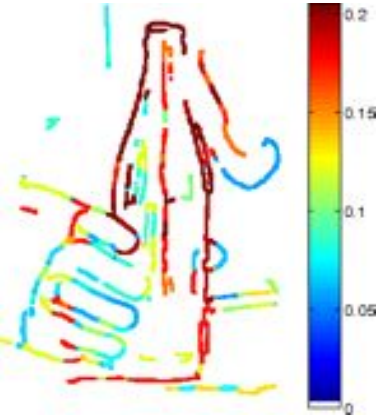
Shape-Context Only





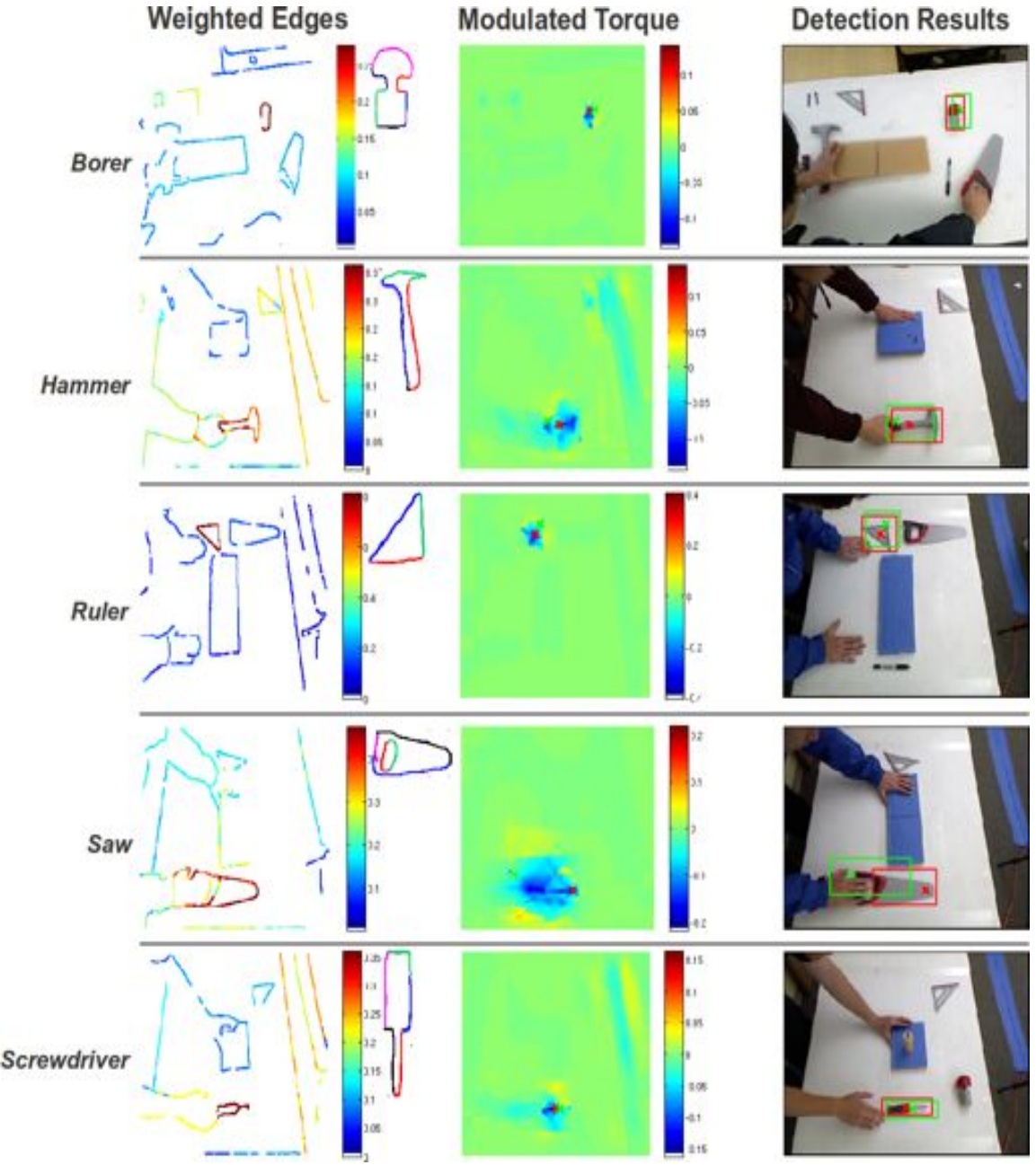


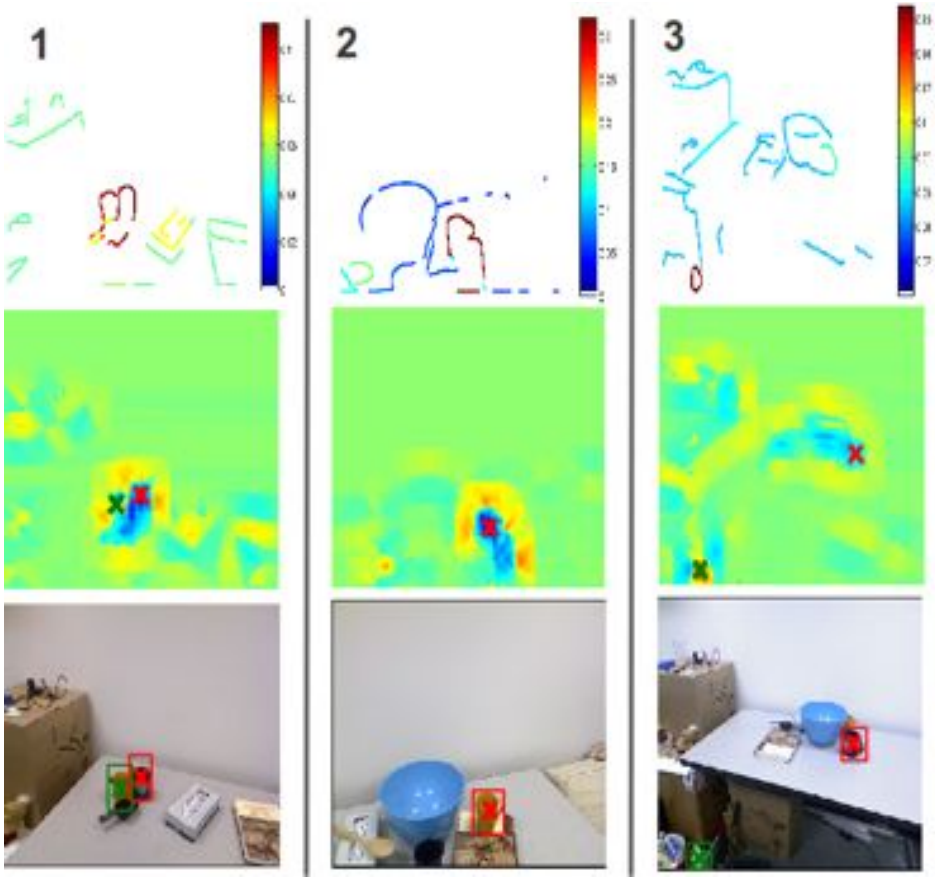
Torque Shape-Context



$$\tau_P^m = \frac{1}{2|P|} \sum_{q \in E(P)} m_{\mathcal{O}}(\tau_{pq})$$

Results

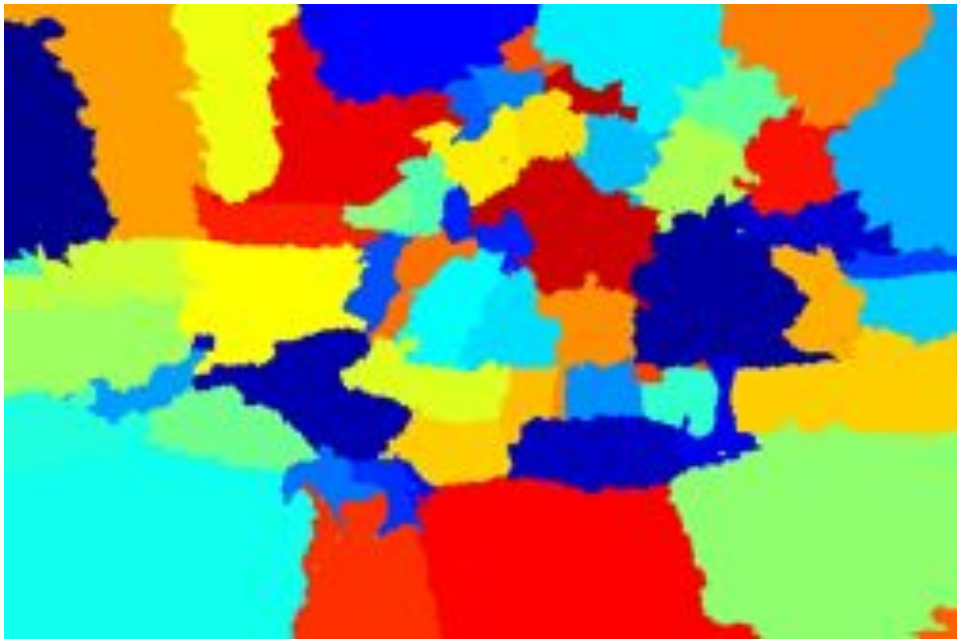

Weighted Edges
Modulated Torque
Detection Results
Borer
Hammer
Ruler
Saw
Screwdriver


1
2
3

Segment the scene

Segmentation with Normalized Cut

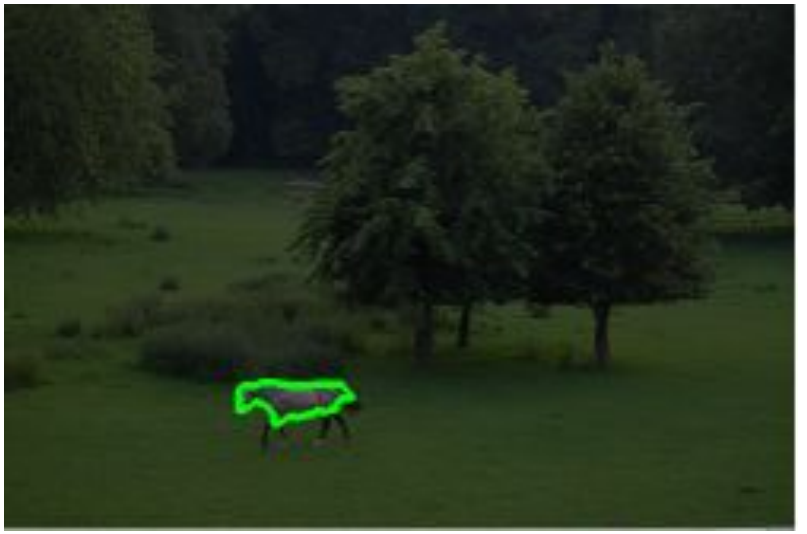
Number of regions = 10

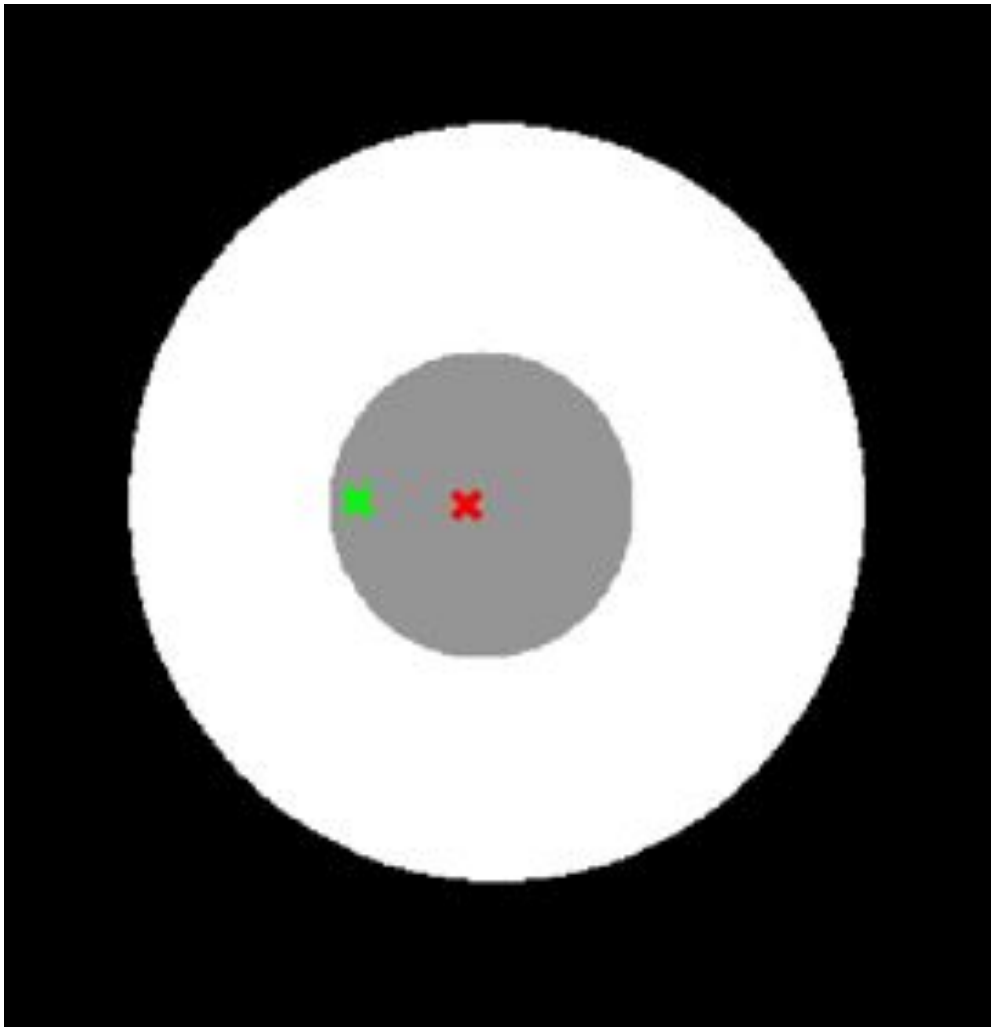


Number of regions = 60

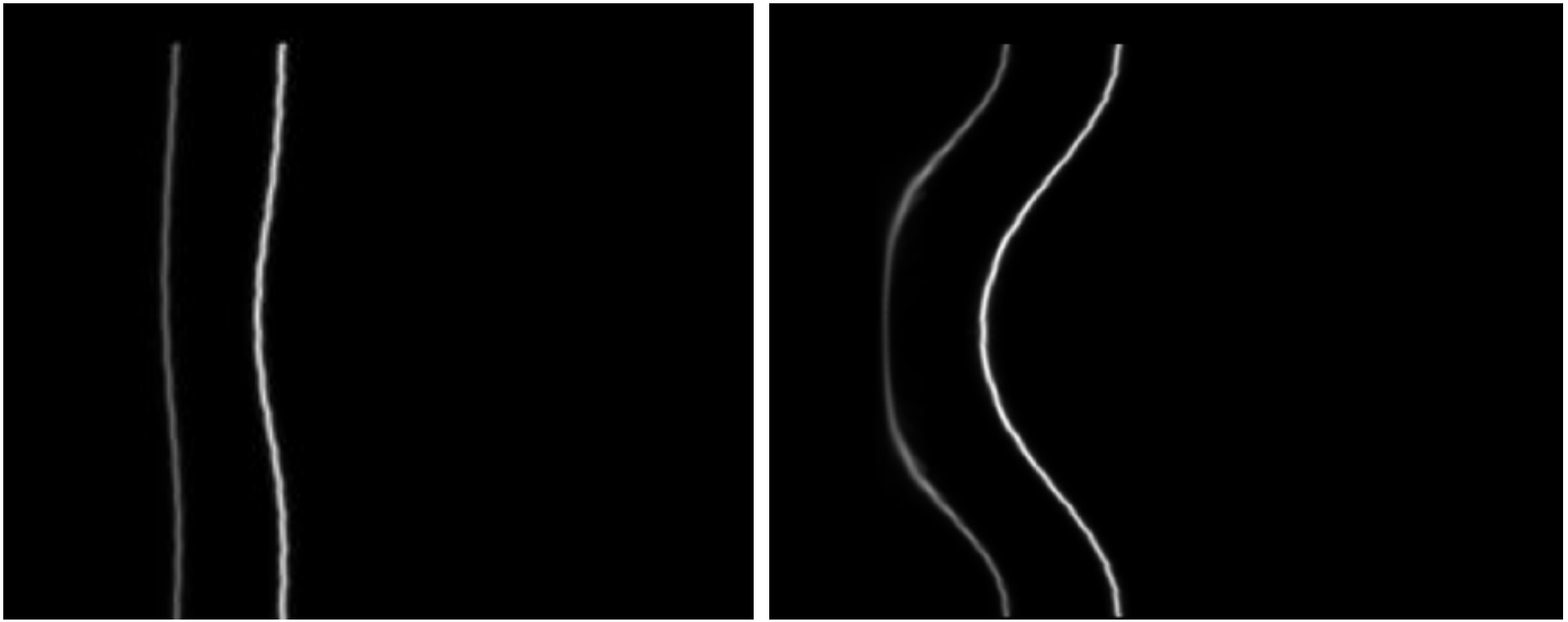
- **Which is the appropriate segmentation?**
  - **Left (trees), Right (horse)**

Segment what you look at

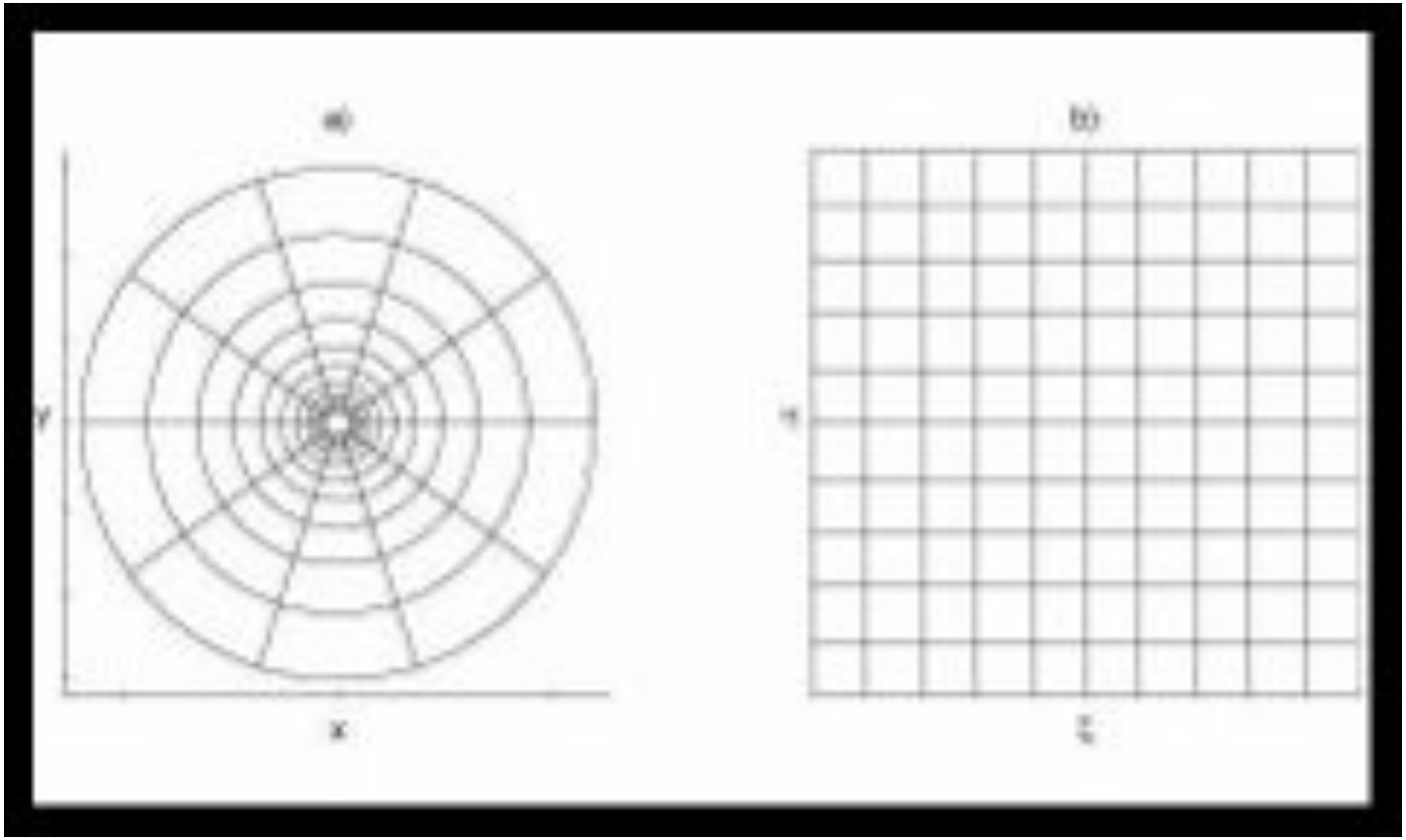


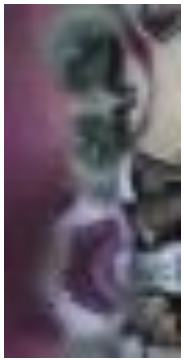


Internal contour: 0.39; outside contour: 0.78
Internal contour length: 100 pixels
Outside contour length: 400 pixels
Cost of tracing each contour
88 = (400 (1 – 0.78)) and 61 = (100 (1 – 0.39))
Smaller (dimmer) contour costs less



Internal contour brightness: 0.39; outside contour: 0.78
Internal contour length: 361 pixels
Outside contour length: 365 pixels
Cost of tracing each contour
80.3 = 365 (1 – 0.78) and 220.21 = 361 (1 – 0.39)
Brighter contour costs less


a)
b)
y
x



Color and texture cues find all edge pixels

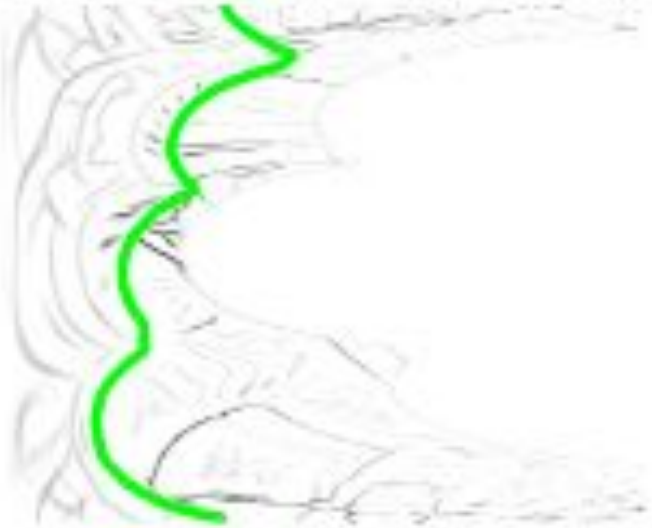
Motion/stereo and surface cues identifies object boundaries

Probabilistic boundary depth map

Segmenting a fixated object

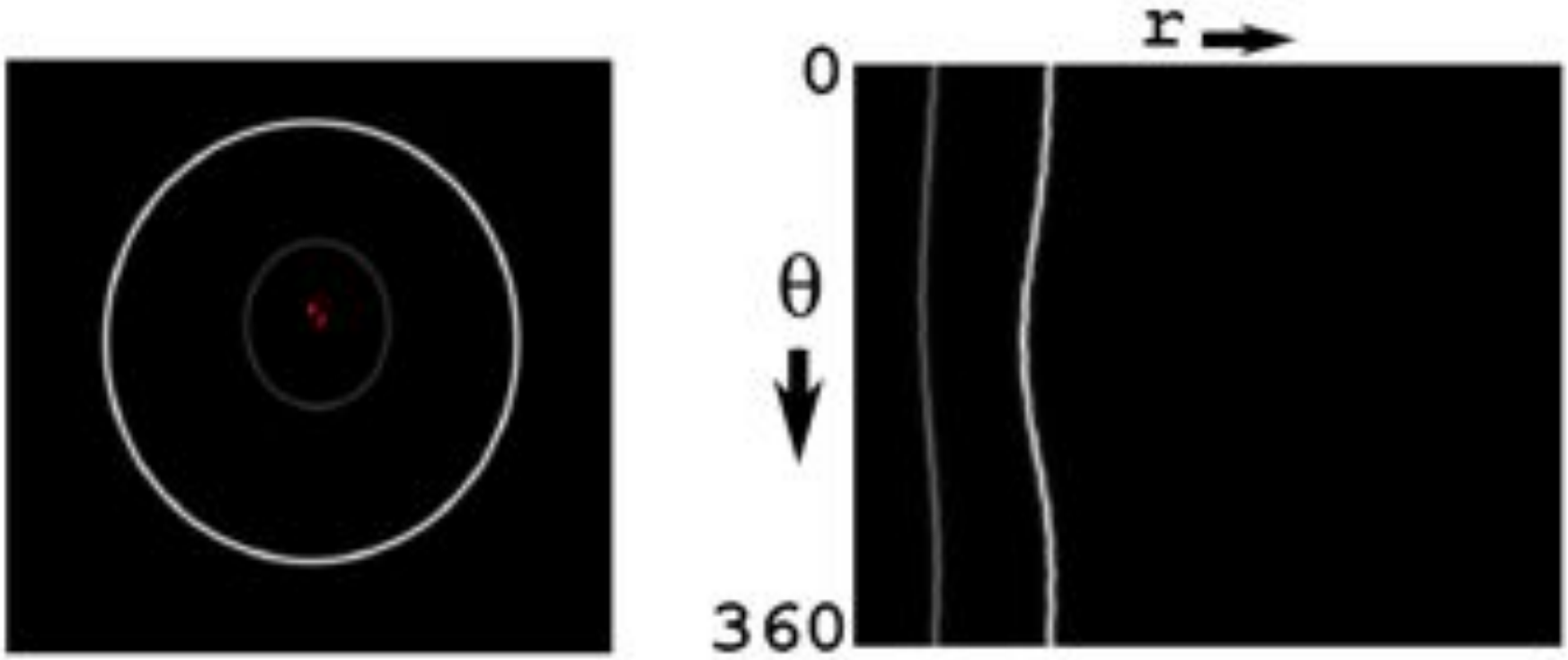
Step 1: Cartesian to polar with fixation as the pole

Step 2: Find the optimal cut through the polar map



Step 1: Cartesian to polar conversion

Scale normalization

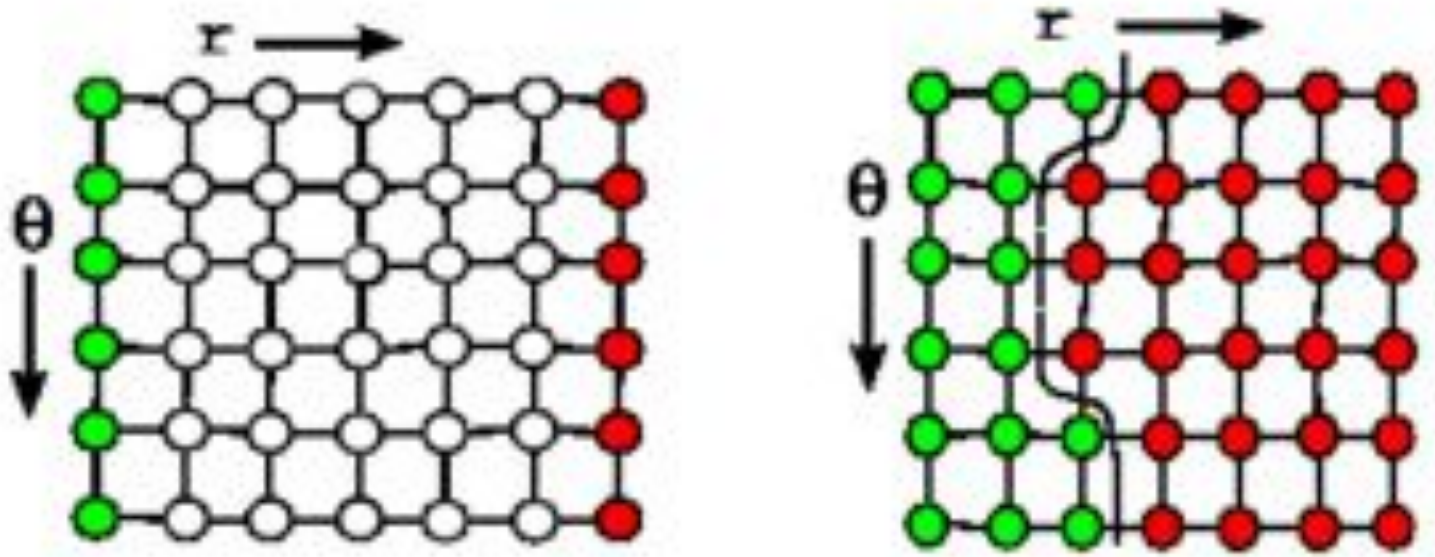


Step 1: Cartesian to polar conversion

Scale normalization

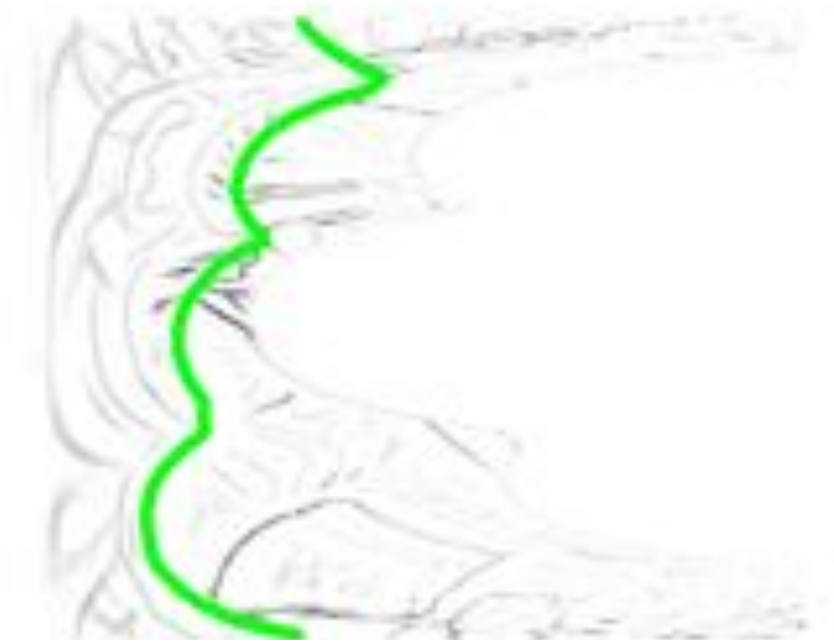
Step 2: Find the optimal cut

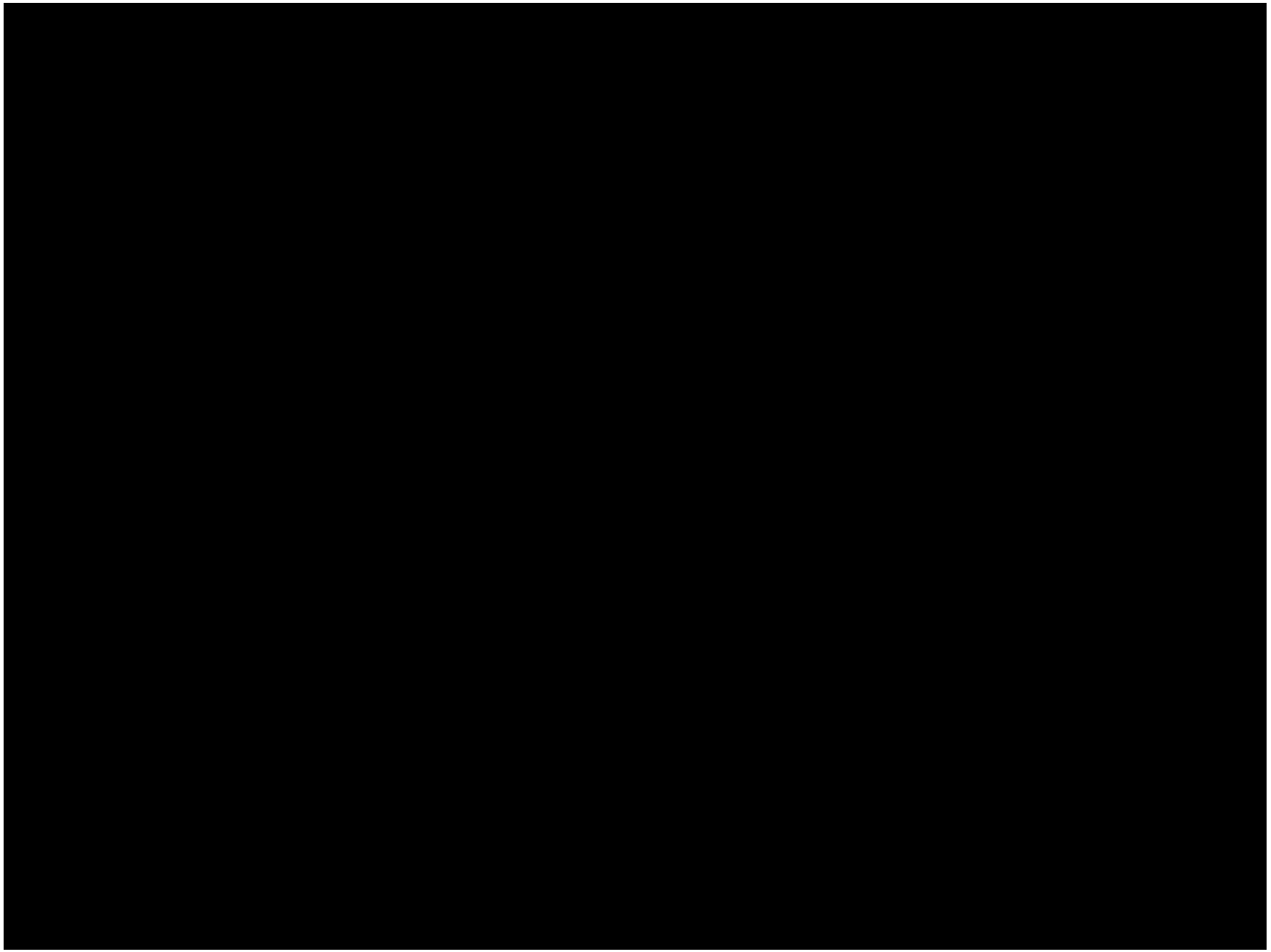
binary labeling problem: Inside Vs outside



Step 2: Find the optimal cut

binary labeling problem: Inside Vs outside





Definition of SIX Primitive Action Consequences

(1)ASSEMBLE: Two objects merge into one object, or two objects build a attachment between them;

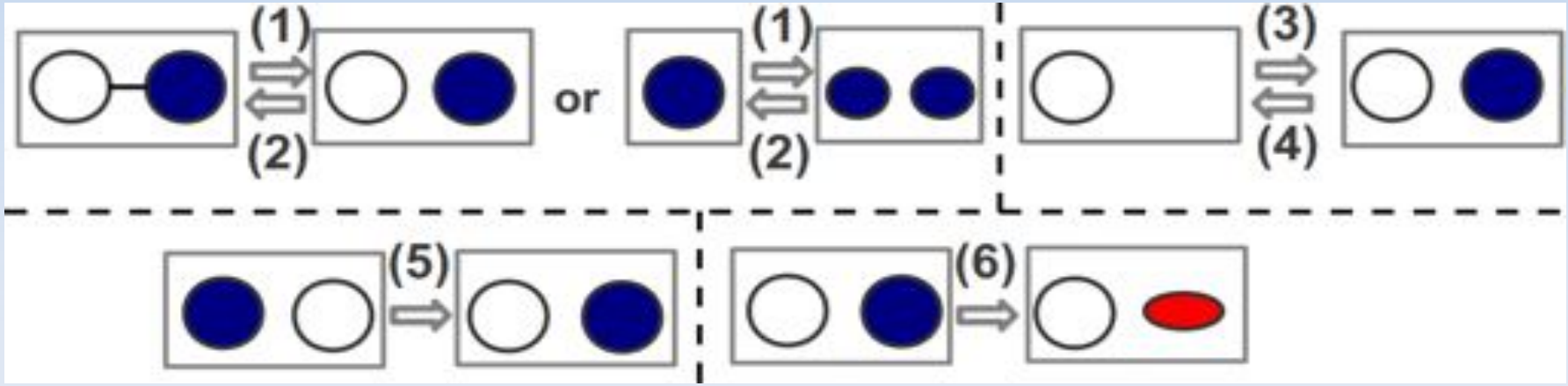
(2)DIVIDE: One object breaks into two objects, or two attached objects breaks the attachment;

(3)CONSUME: An object is disappeared from the visual space;

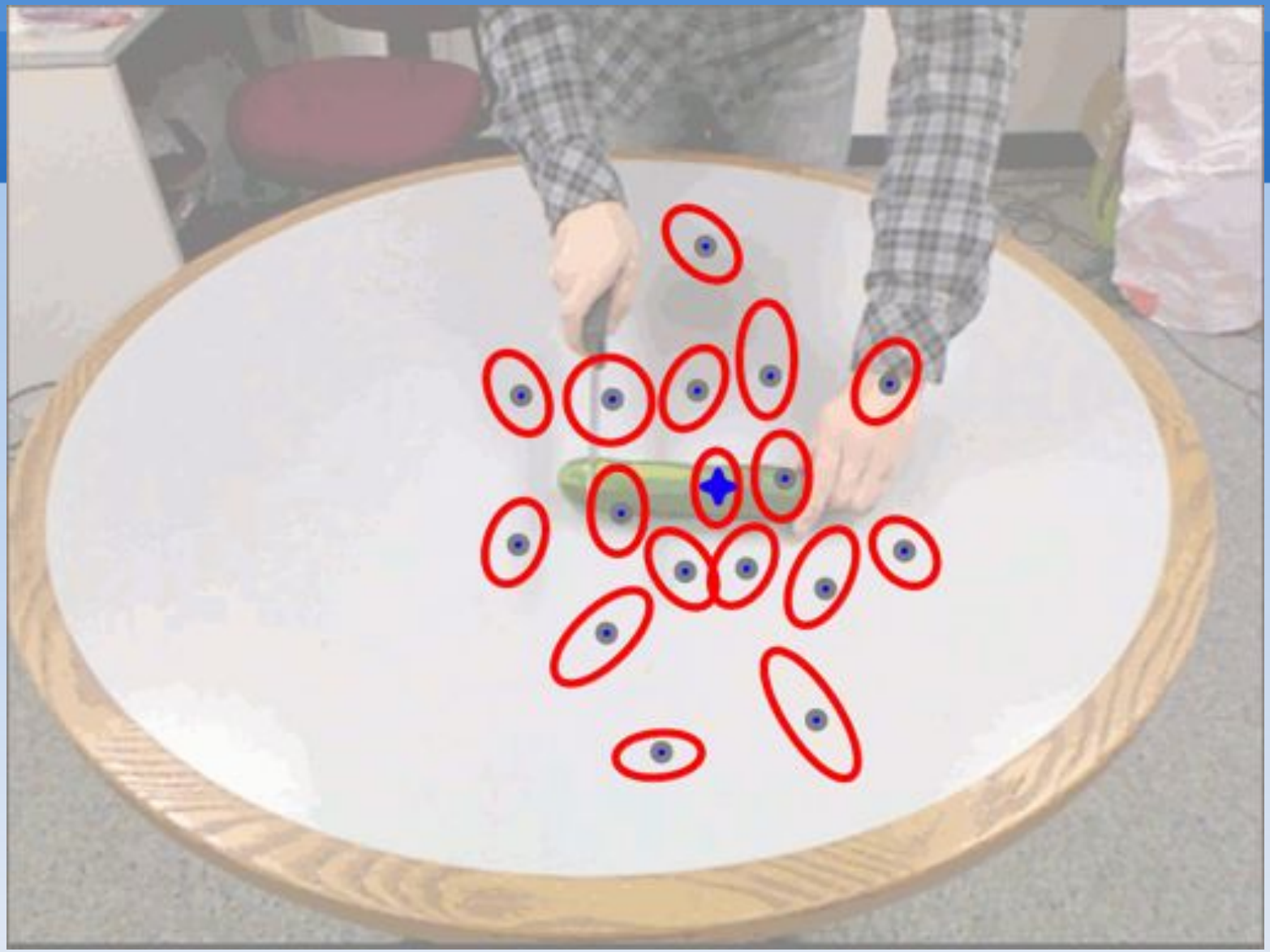
(4)CREATE: An object is brought to, or emerge into the visual space;

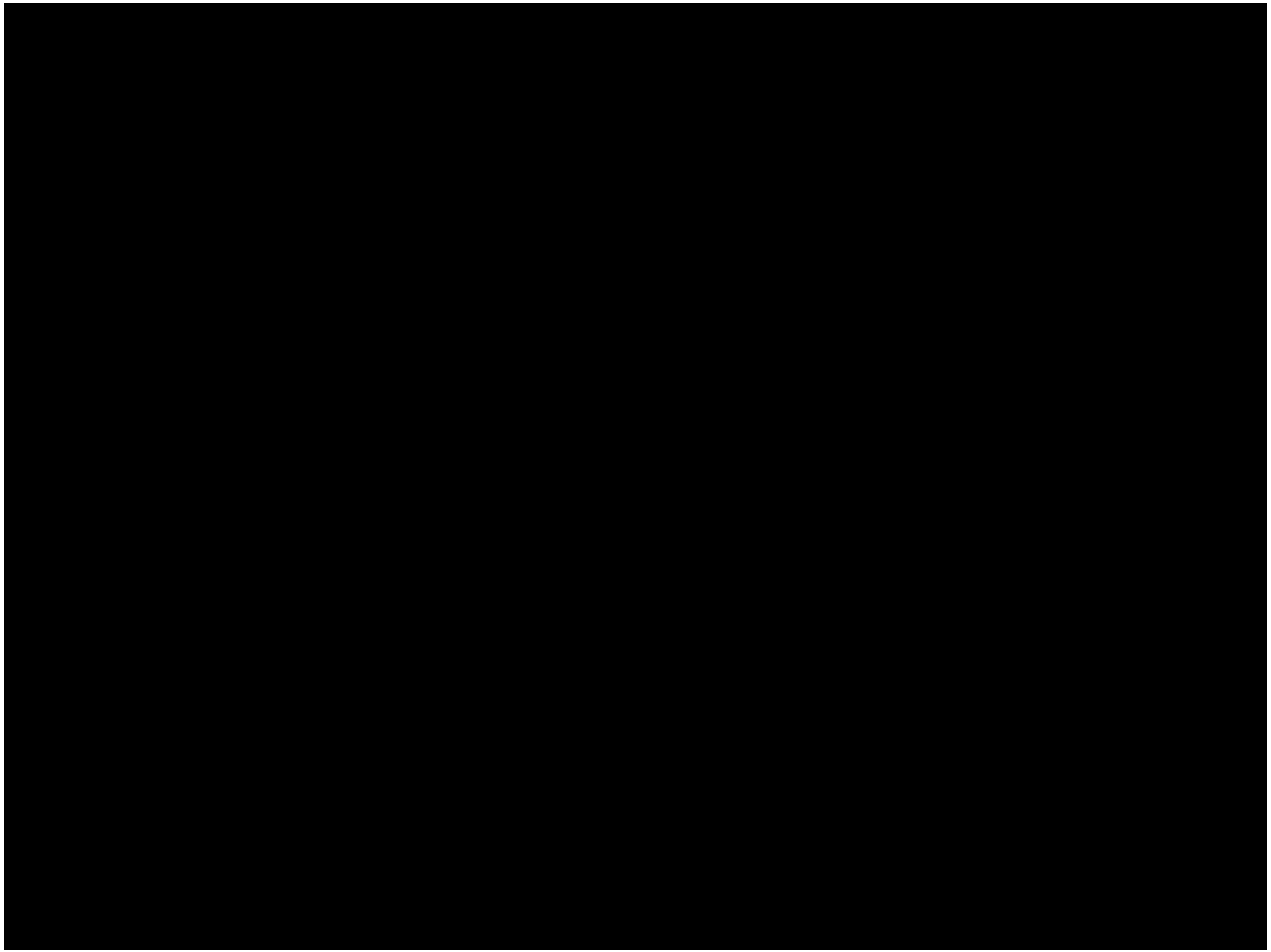
(5)TRANSFER: An object is moved from one location to another location;

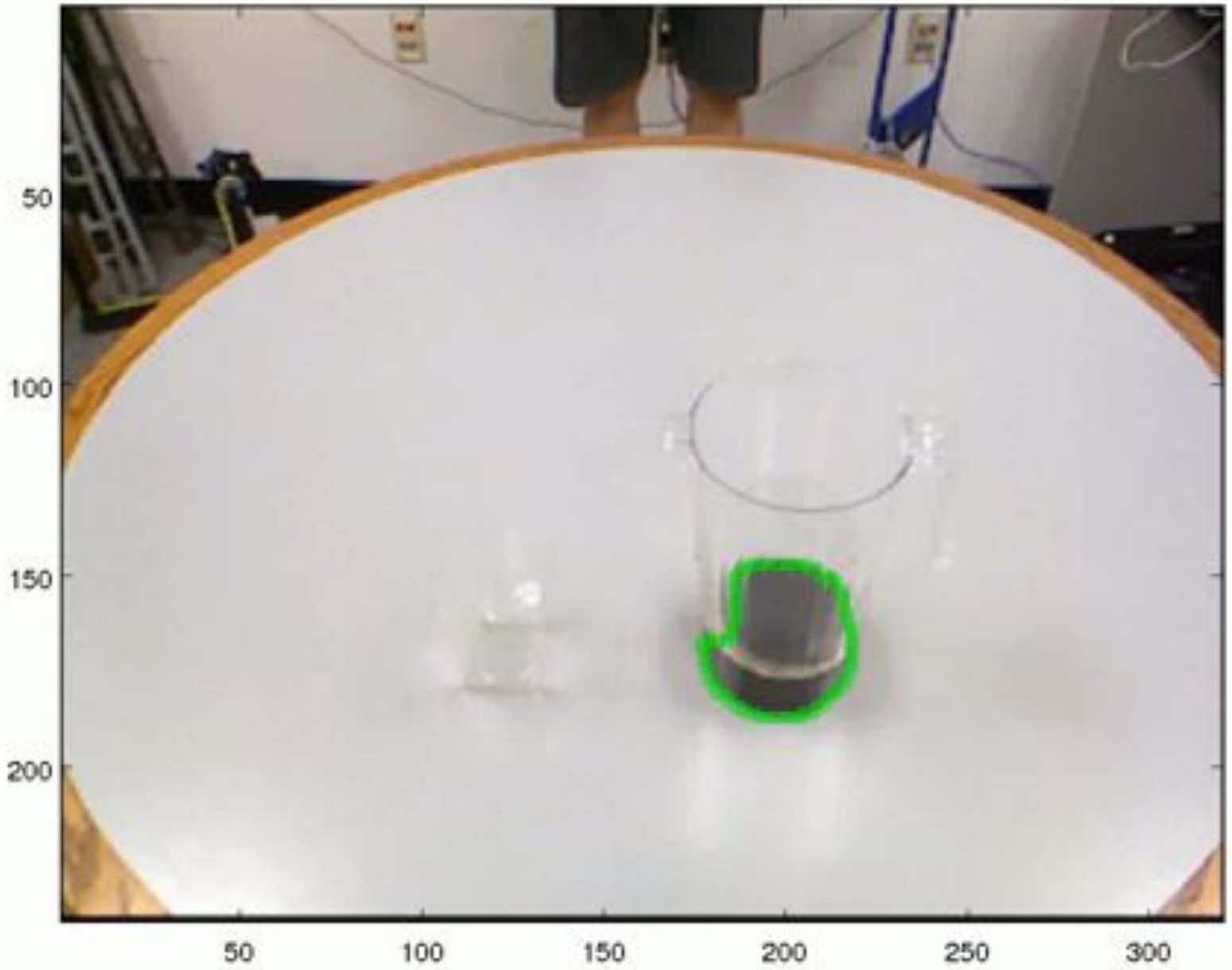
(6)DEFORM: An object has an appearance change.



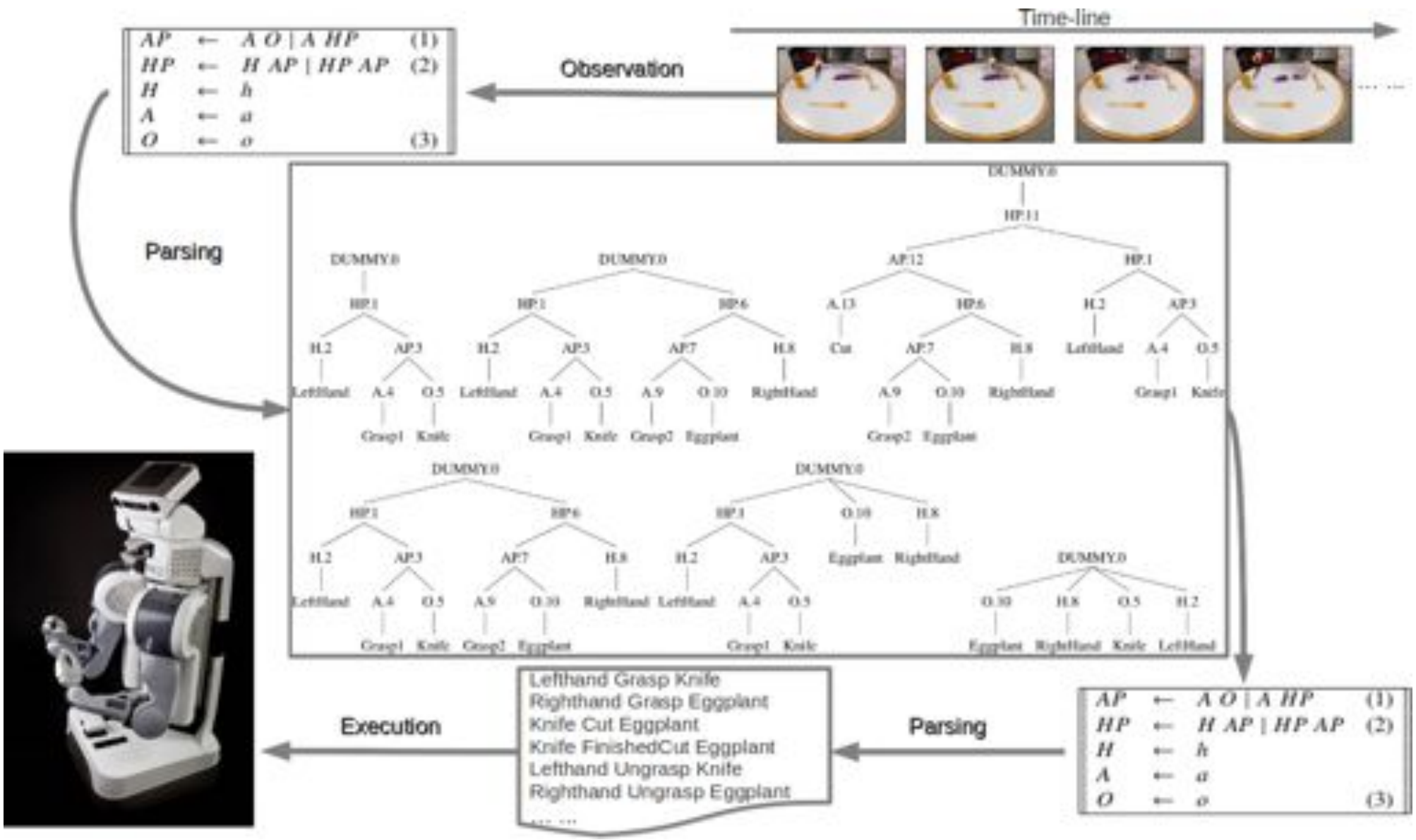
Action Consequences Detection Through Object Monitoring

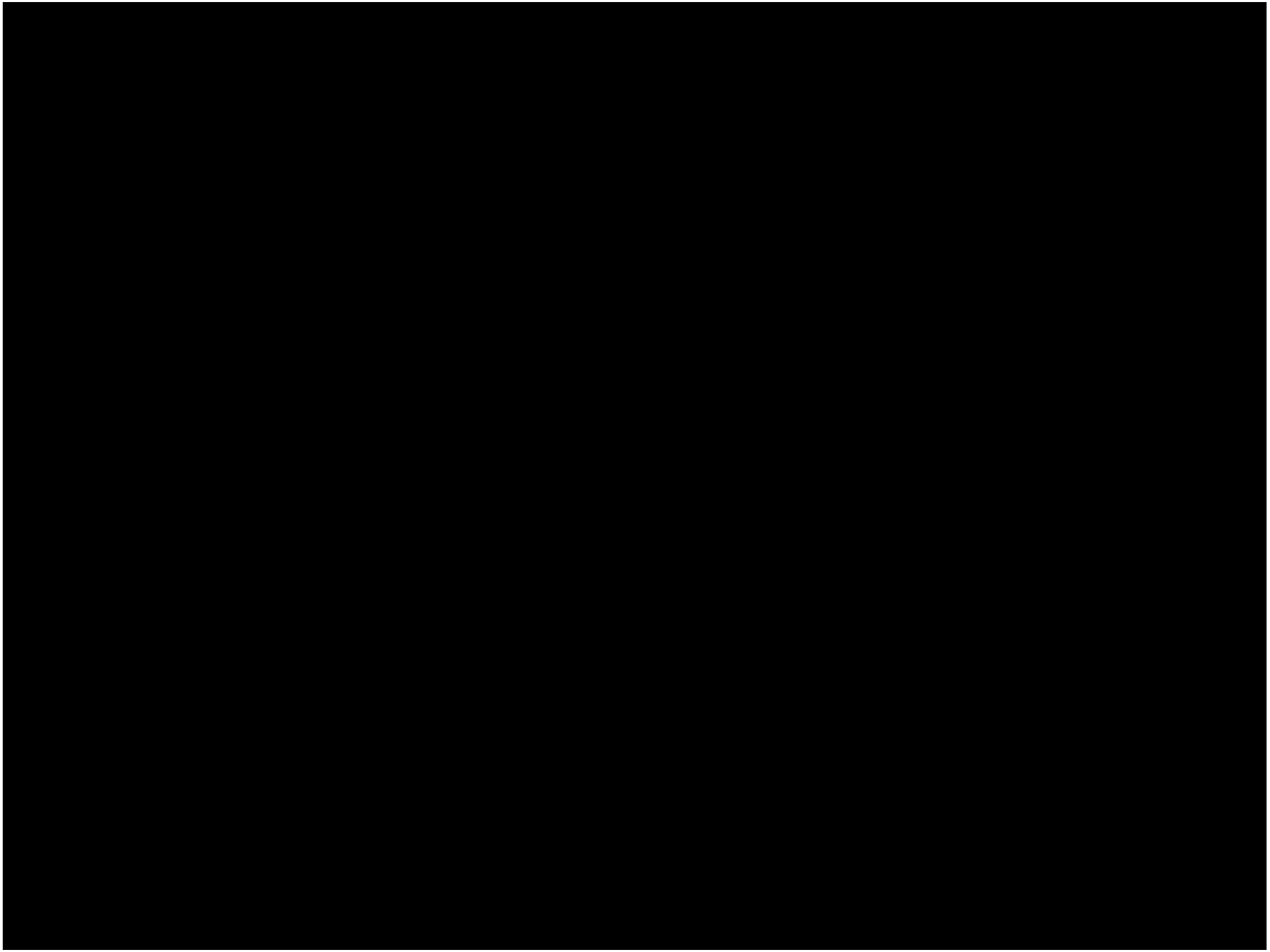




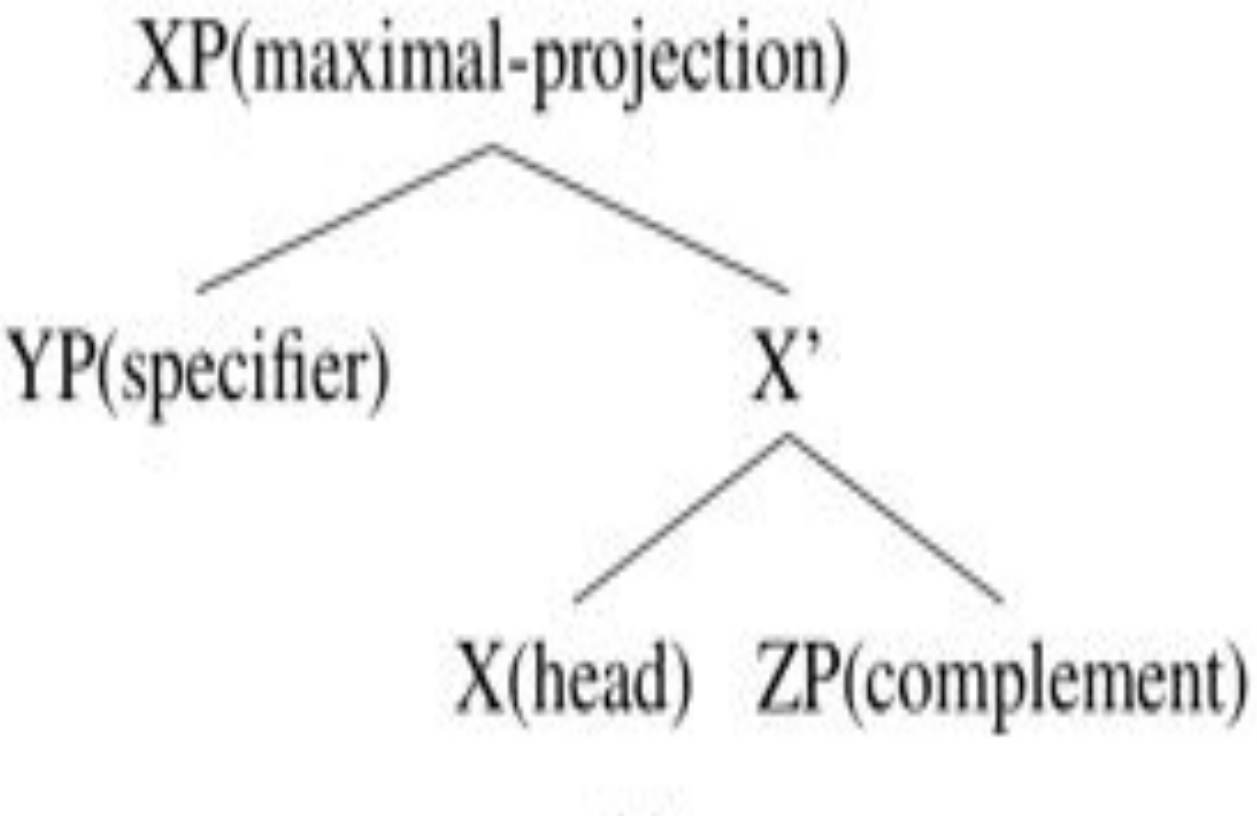

50
100
150
200
50
100
150
200
250
300

An Example for Cognitive Robots

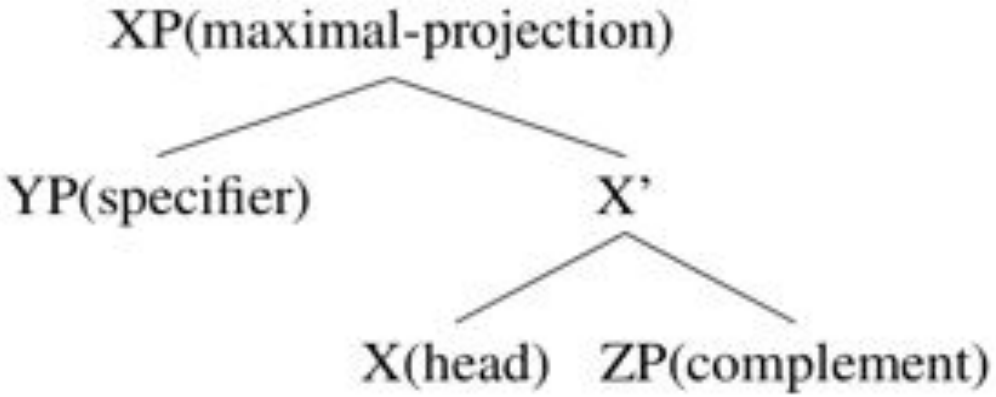




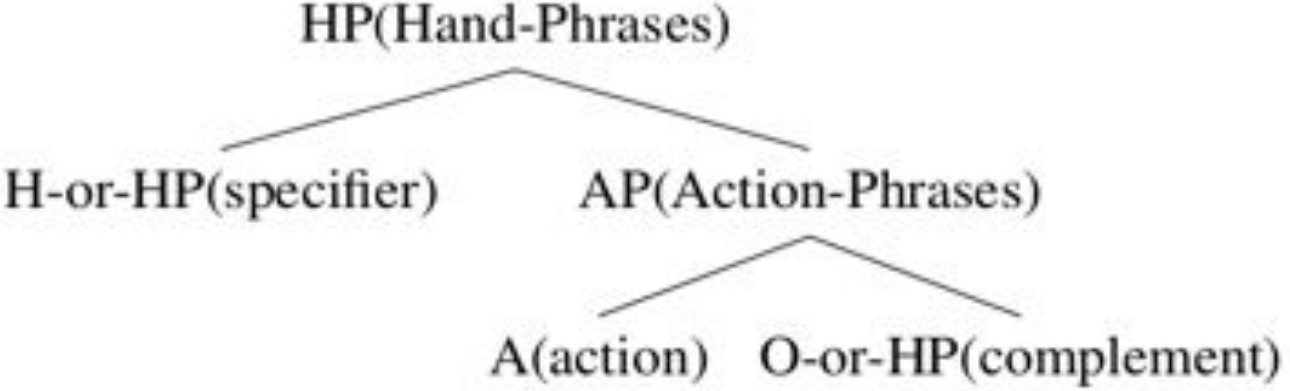
X-bar Schema




IP
NP
I
N'
I [–Pst]
VP
N
V
He
V
PP
V
NP
P
studies
N
P
NP
N
at
Det
N
linguistics
the
N
university


XP(maximal-projection)
YP(specifier)
X'
X(head)
ZP(complement)

(a)


HP(Hand-Phrases)
H-or-HP(specifier)
AP(Action-Phrases)
A(action)
O-or-HP(complement)

(b)

The Grammar Rules: terminal symbols

$$
\begin{aligned}
H &\leftarrow h \\
A &\leftarrow a \\
O &\leftarrow o
\end{aligned}
$$

H for Hands, A for Actions, O for Objects. AP is Action Phrases, HP is Hand Phrases

The Grammar Rules: Rules to make Action Phrases (AP)

$$AP \leftarrow A\ O \mid A\ HP$$

Insight behind:

– An “Action” (A) can be applied to an “Object” (O) directly,

– or to a “Hand Phrase” (HP), which in turn contains an “Object” (O).

The Grammar Rules: Rules to make Hand Phrases (HP)

$$HP \leftarrow H\ AP \mid HP\ AP$$

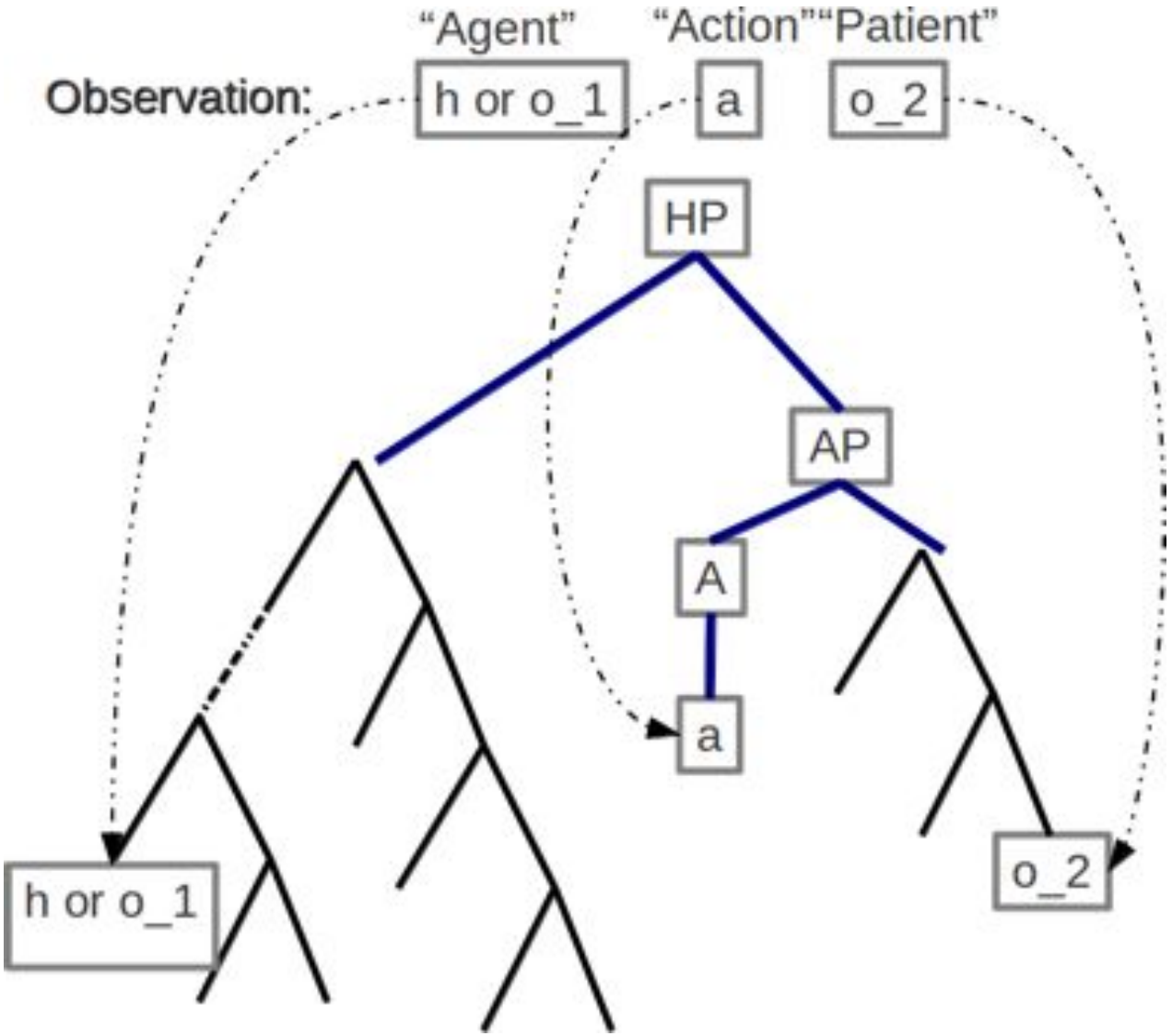
Insight behind:

– An “Action Phrase” (AP) can be combined either with the “Hand”(H), or a “Hand Phrase”, which recursively builds up the “Hand Phrase”.

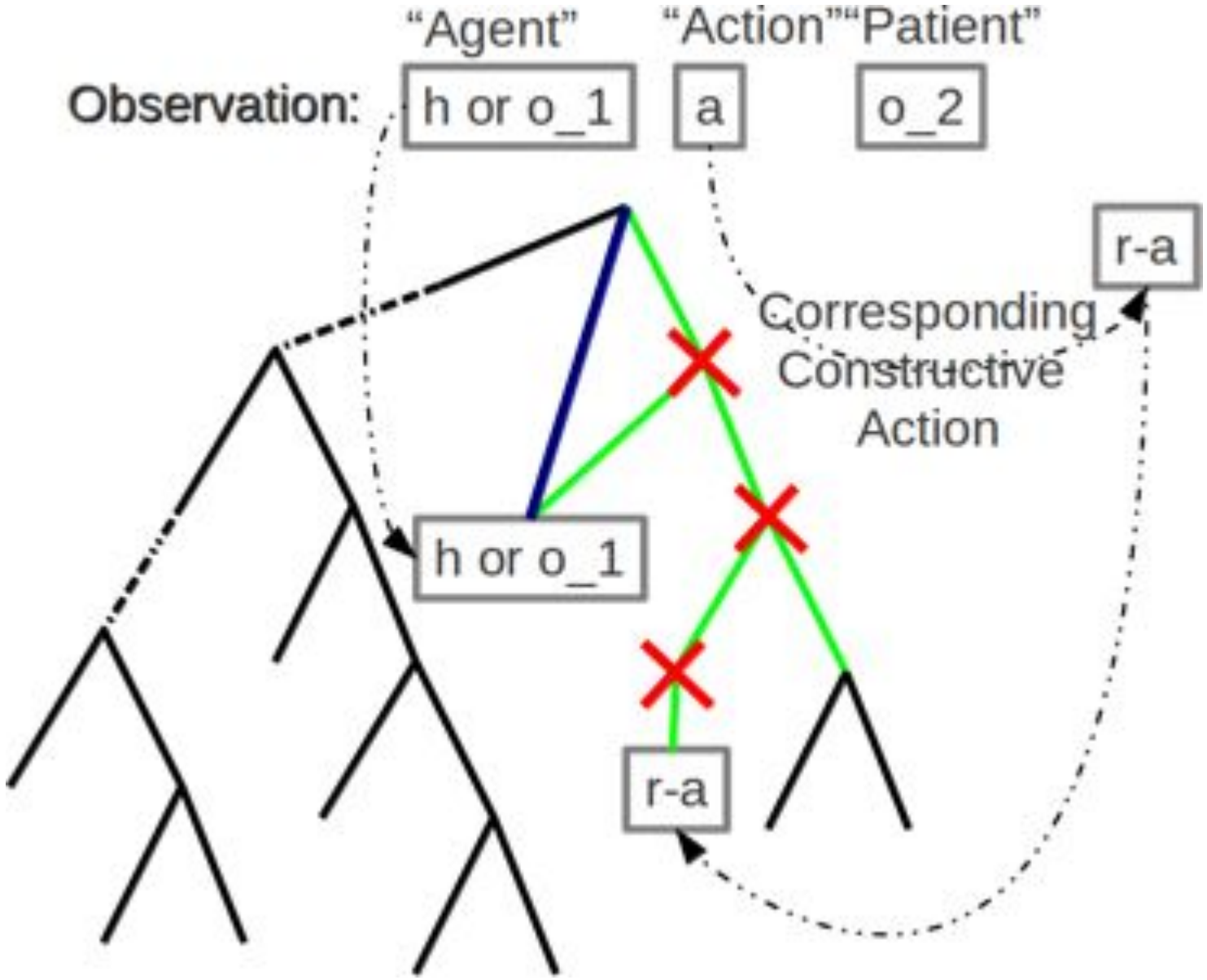
How to parse?

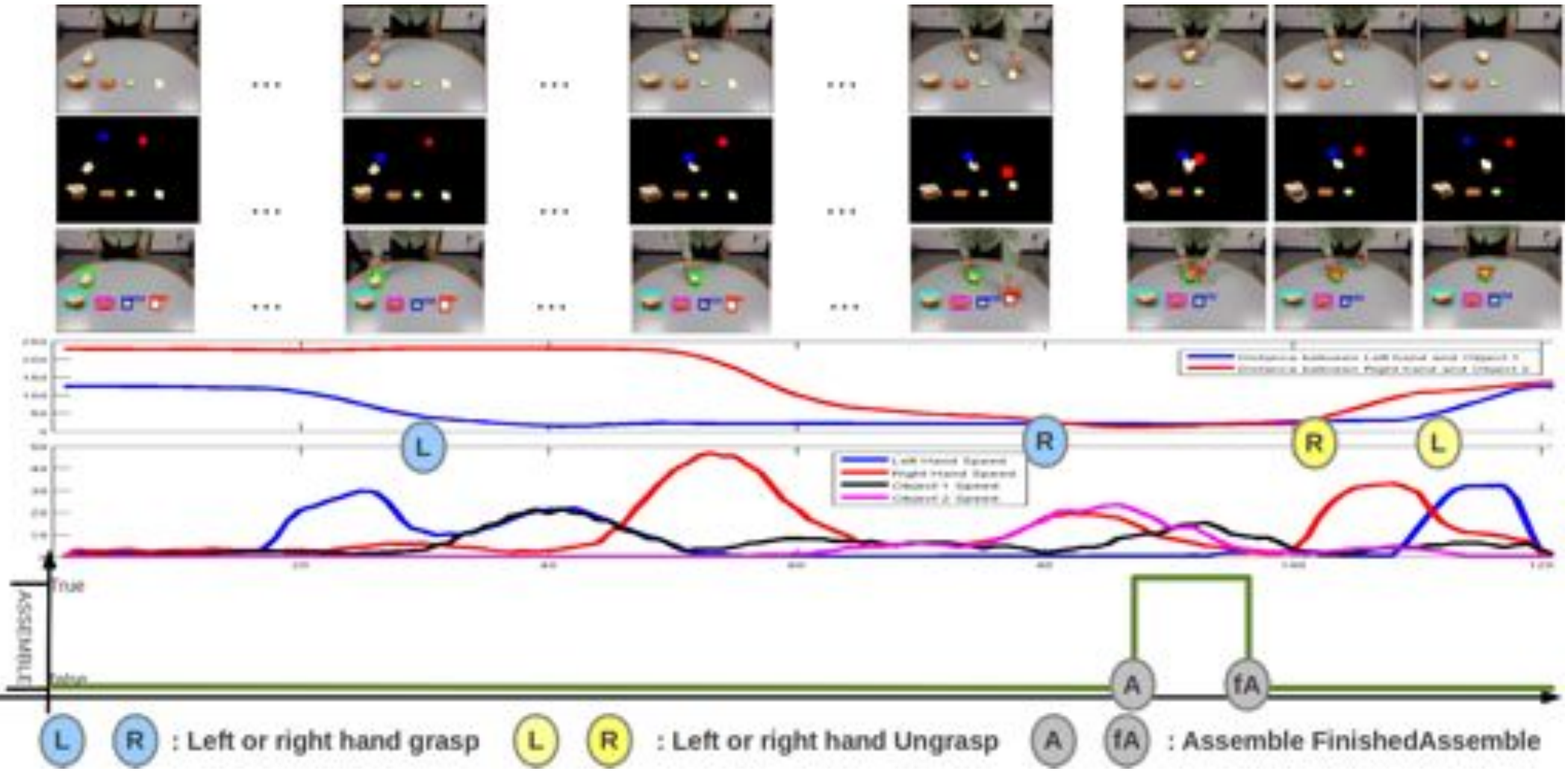
- The visual observations are obtained in a temporal sequence;
- A parsing algorithm is needed for the grammar should to dynamically update the semantic tree structures on the fly;
- Constructive (“Grasp”, “Cut”, etc) and Destructive Actions (“unGrasp”, “finishedCut”, etc).

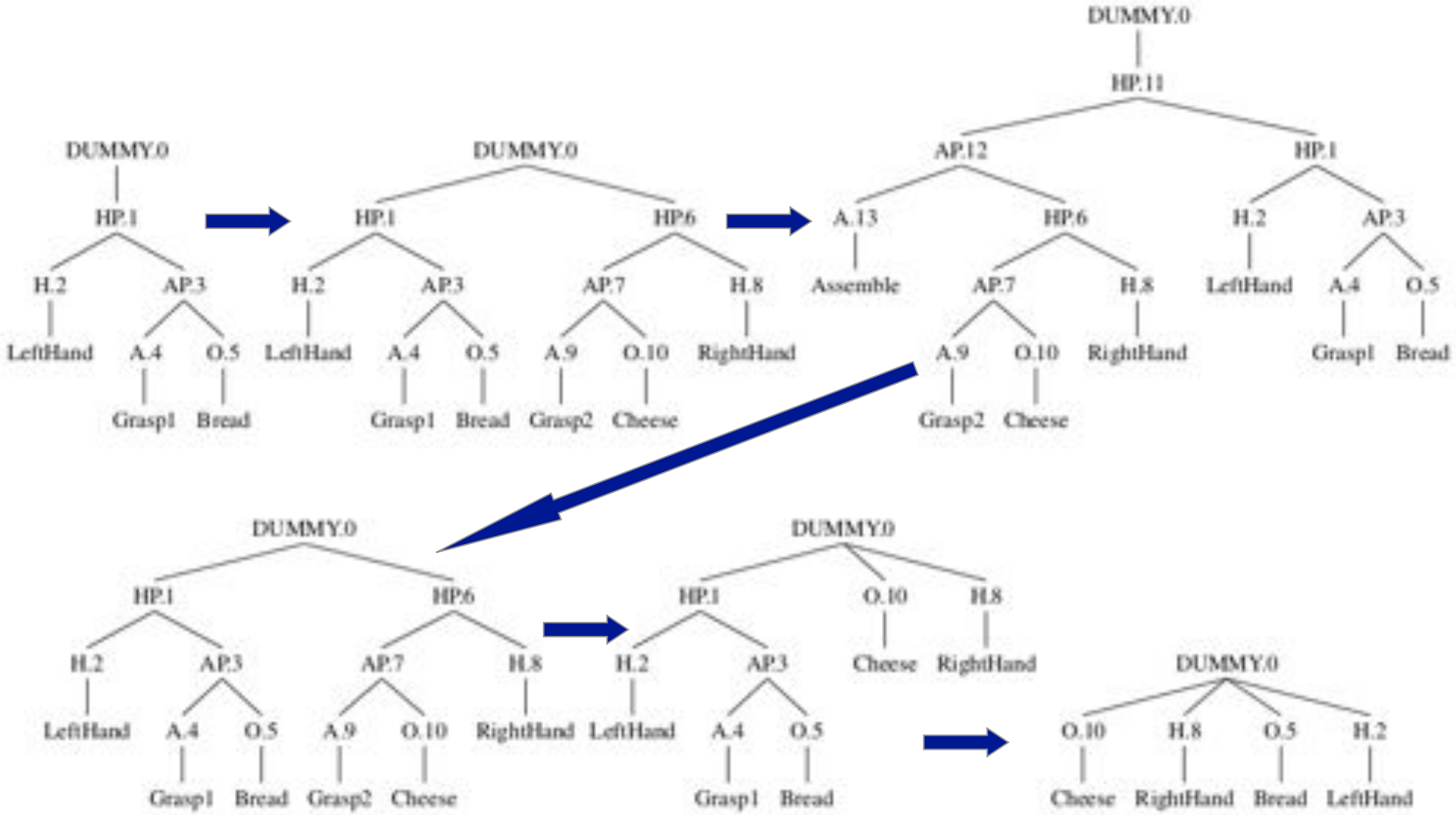
The Parsing Algorithms: Construction

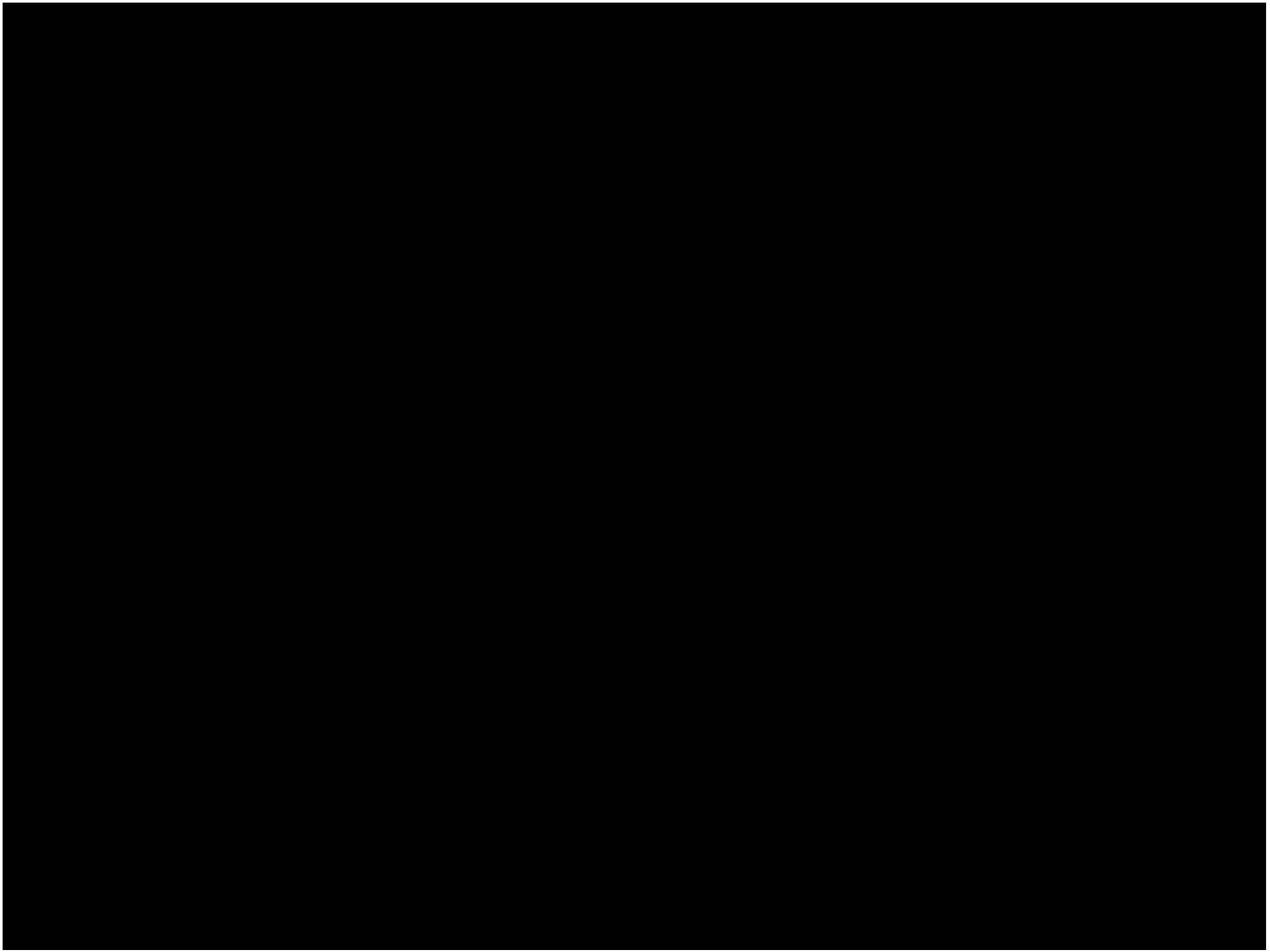


The Parsing Algorithms: Destruction



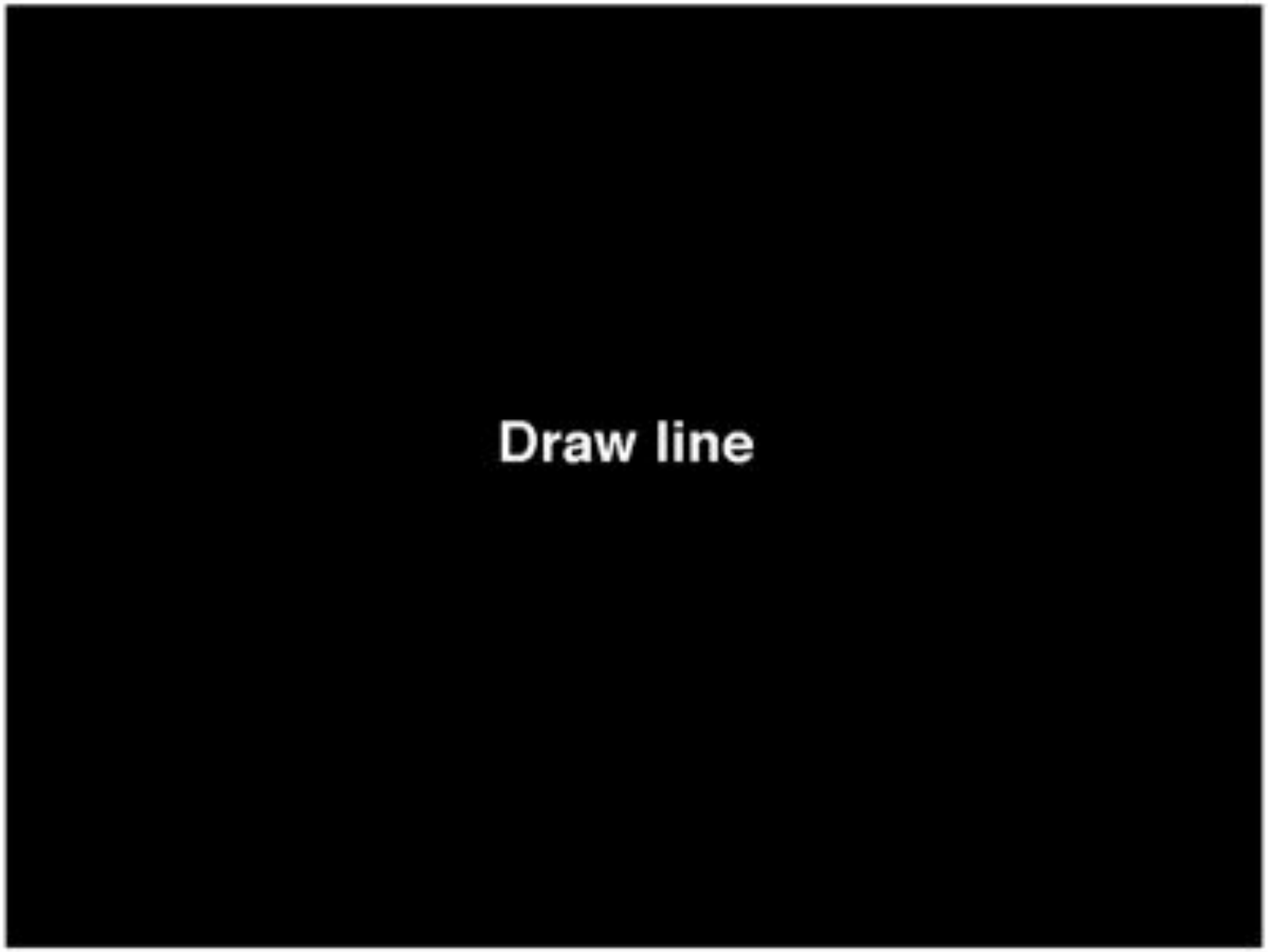

L
R
R
L
Assemble
True
False
A
fA
L
R
: Left or right hand grasp
L
R
: Left or right hand Ungrasp
A
fA
: Assemble FinishedAssemble


DUMMY.0
HP.1
H.2
AP.3
LeftHand
A.4
O.5
Grasp1
Bread
DUMMY.0
HP.1
HP.6
H.2
AP.3
AP.7
H.8
LeftHand
A.4
O.5
A.9
O.10
RightHand
Grasp1
Bread
Grasp2
Cheese
DUMMY.0
HP.11
AP.12
HP.1
A.13
HP.6
H.2
AP.3
Assemble
AP.7
H.8
LeftHand
A.4
O.5
A.9
O.10
RightHand
Grasp1
Bread
Grasp2
Cheese
DUMMY.0
HP.1
HP.6
H.2
AP.3
AP.7
H.8
LeftHand
A.4
O.5
A.9
O.10
RightHand
Grasp1
Bread
Grasp2
Cheese
DUMMY.0
HP.1
O.10
H.8
H.2
AP.3
Cheese
RightHand
LeftHand
A.4
O.5
Grasp1
Bread
DUMMY.0
O.10
H.8
O.5
H.2
Cheese
RightHand
Bread
LeftHand



What's next ?

- How to do the learning?
  - From various observations.
  - From raw language resources like recipes or wikipedia.
  - From human instructions.
- Human Robot Interaction and Collaboration?
  - Action prediction.
  - Warning for mistakes.
- Primitive action recognition, beyond hand trajectories?


Draw line

UNIVERSITY OF MARYLAND 18 56

NATIONAL INSTITUTES OF HEALTH

poeticon++

NSF

QUALCOMM®