Towards the Computational Perception of Action

Richard Mann – NEC Research Institute, Princeton, NJ.
Allan Jepson – Department of Computer Science, University of Toronto.

1. THE PROBLEM

1.1 Interpreting action among objects

Sample input sequence (hand lifts can)

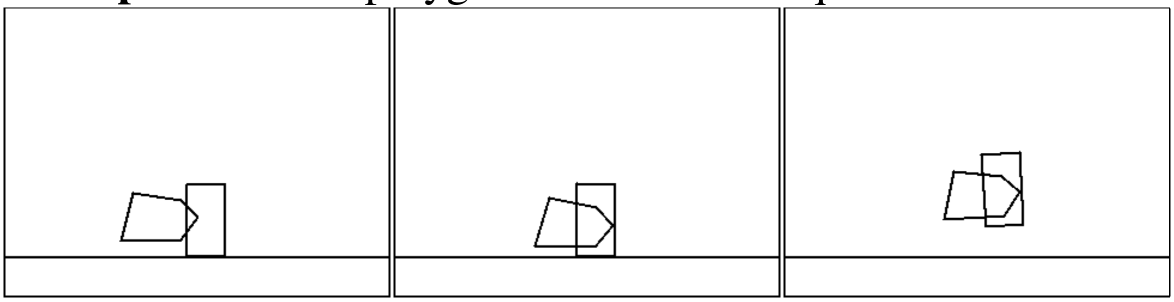
Questions:

- What *action(s)* occur in sequence (e.g., lift, drop, hit, etc.)
- What are the *roles* of the various objects in the action.

1.2 Motivation

Actions can be recognized *without* object recognition or domain-specific rules.

Example: Tracked polygons from above sequence



Implicit inferences:

- Can is ‘passive object’ (supported by table).
- Hand is ‘active object’ (moving above table).
- Hand lifts can off table (by attaching to it).

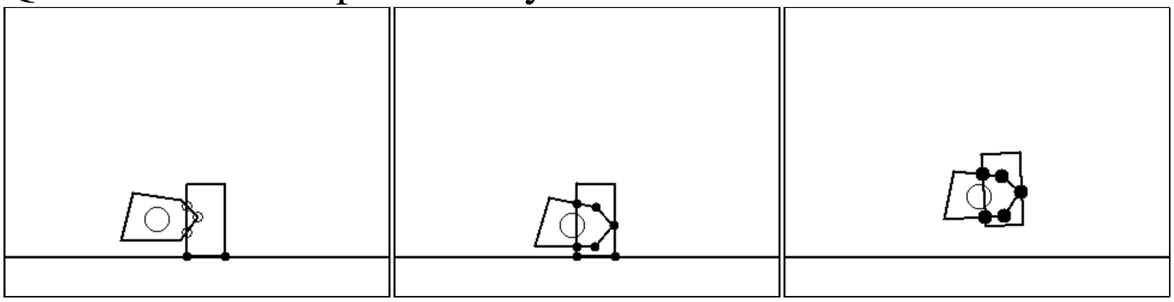
1.3 Approach

Analyse *Newtonian mechanics* of a simplified scene model.

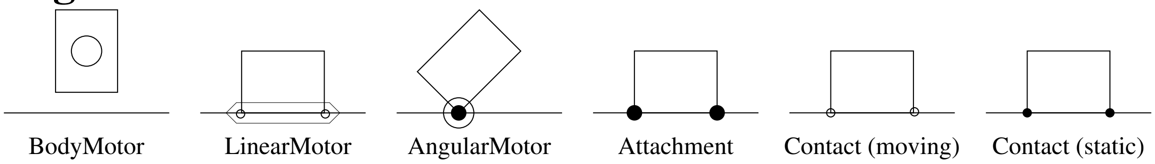
- 2D polygonal objects (layered depth model),
- Gravity and ground plane,
- Contact and attachment between objects,
- 'Active' vs. 'passive' objects.

1.4 Output of System

Qualitative Description of Dynamics



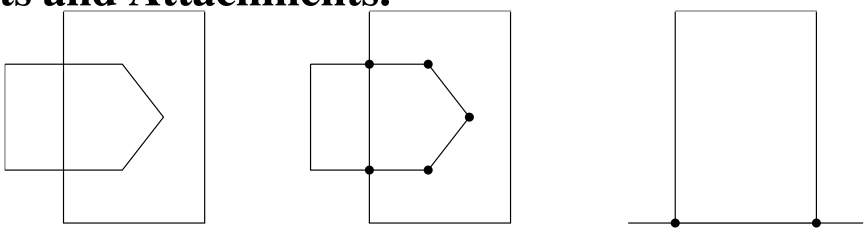
Legend:



2. UNDERSTANDING SCENE DYNAMICS
(Mann, Jepson, & Siskind, 1996)

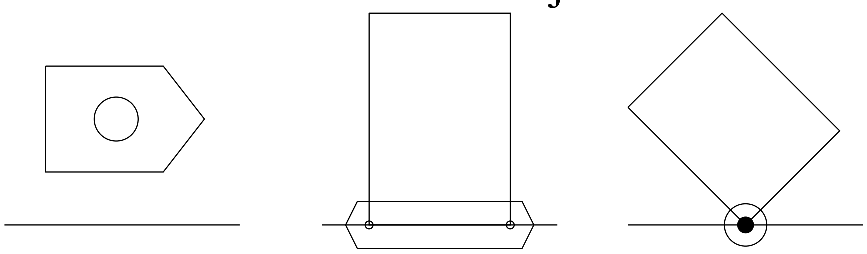
2.1 Representation

- **Objects.** 2D rigid polygons; “Layered” depth model.
- **Contacts and Attachments.**



Disjoint | Overlap | Abut

- **Masses.** Mass, mass distribution (inertia).
- **Forces.** Gravity, contact forces (normal force, frictional force).
- **Ground plane.** Used for table top.
- **Motors.** Assertions about “active” objects.



BodyMotor | LinearMotor | AngularMotor

2.2 Feasible Interpretations

Definition: Interpretation is *feasible* if observed motion is consistent with physics-based model.

Testing Feasibility by 'Force-balancing'

- For each object, sum forces due to gravity, contact with other objects, and forces due to motors.
- masses are variable (constrained to be positive).
- solution must explain observed accelerations according to Newtonian mechanics.
- formulate as *constraint satisfaction problem*.
  (Reduces to *linear programming* in our case)

Sample inferences:

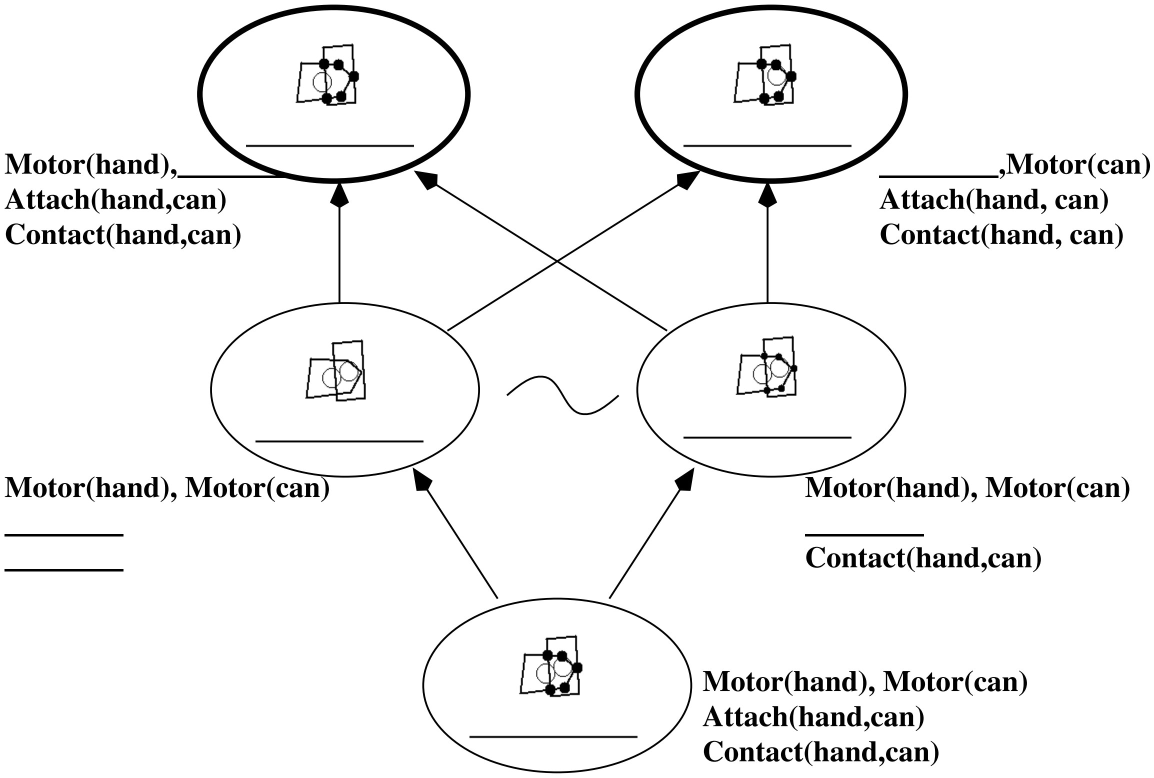
- 'passive' objects: must be *supported* (otherwise *falling*).
- non-attached objects: can 'push' but not 'pull'.

2.3 Preferred Interpretations

Observation: Trivial solutions always exist with motors on all objects!
To get *informative* interpretations, prefer *minimum* sets of motors and attachments.

Example: Preference order on interpretations for lifting phase of `coke` sequence. (Motors minimized at higher priority than attachments; contacts allowed to vary.)

- hand lifts can.
- can lifts hand.



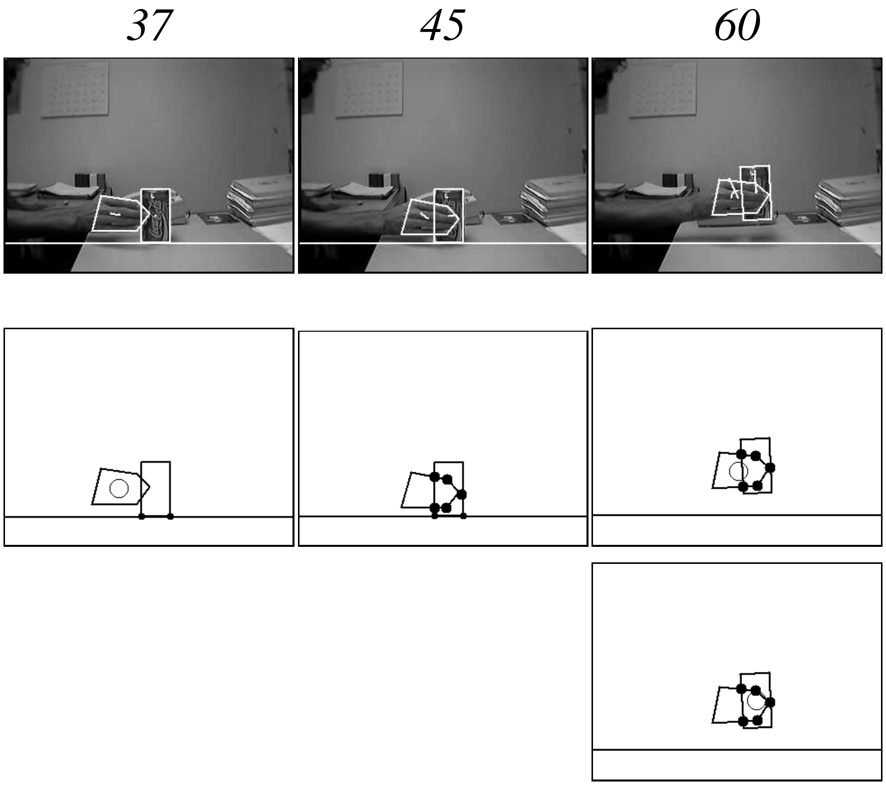
Note: Based on only the instantaneous motion, there are two preferred interpretations!

3. INTEGRATE OVER TIME

3.1 Problems with Instantaneous Analysis

1. **Ambiguity.** When objects interact, we can't tell which object (if any), is generating forces.

2. **Anomalous frames.** Motion discontinuities (collisions, contact changes), or starts and stops of motion.



Frame 45 : *Anomaly:* no motor required; instead the hand is attached to the can.
Frame 60 : *Ambiguity:* Is the motor on the hand or the can?

3.2 Ascribing Properties to Objects

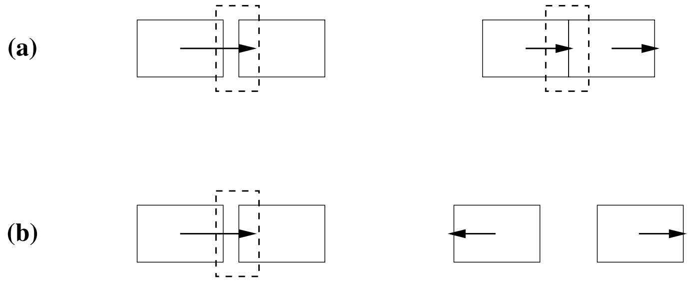
1. Ascribe set of *properties* or *abilities* to objects.
2. Choose minimum set of properties for *whole sequence*.
   (Note: requires correspondence of objects in all frames.)

Representation

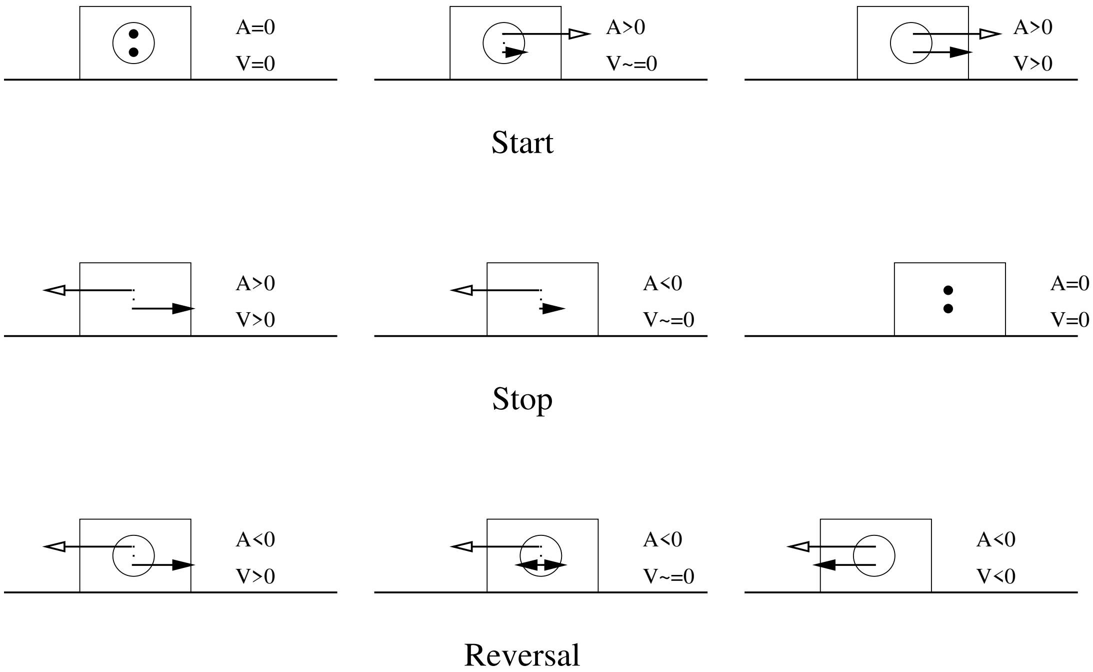
- $\textsc{Flyer}(o)$ — o can generate force and torque on itself;
- $\textsc{Driver}(o)$ — o can exert tangential force along its edges;
- $\textsc{Rotor}(o)$ — o can exert torque at its vertices;
- $\textsc{Grasper}(o)$ — o can attach to other objects.

3.3 Sequence Processing

(i) Discontinuity detection: Collisions detected by impending penetration followed by either abutting contact (a), or separation (b).



(ii) Start-stop detection: Ambiguity occurs when tangential velocity is so small that the sign cannot be reliably determined.



4. EXPERIMENTS

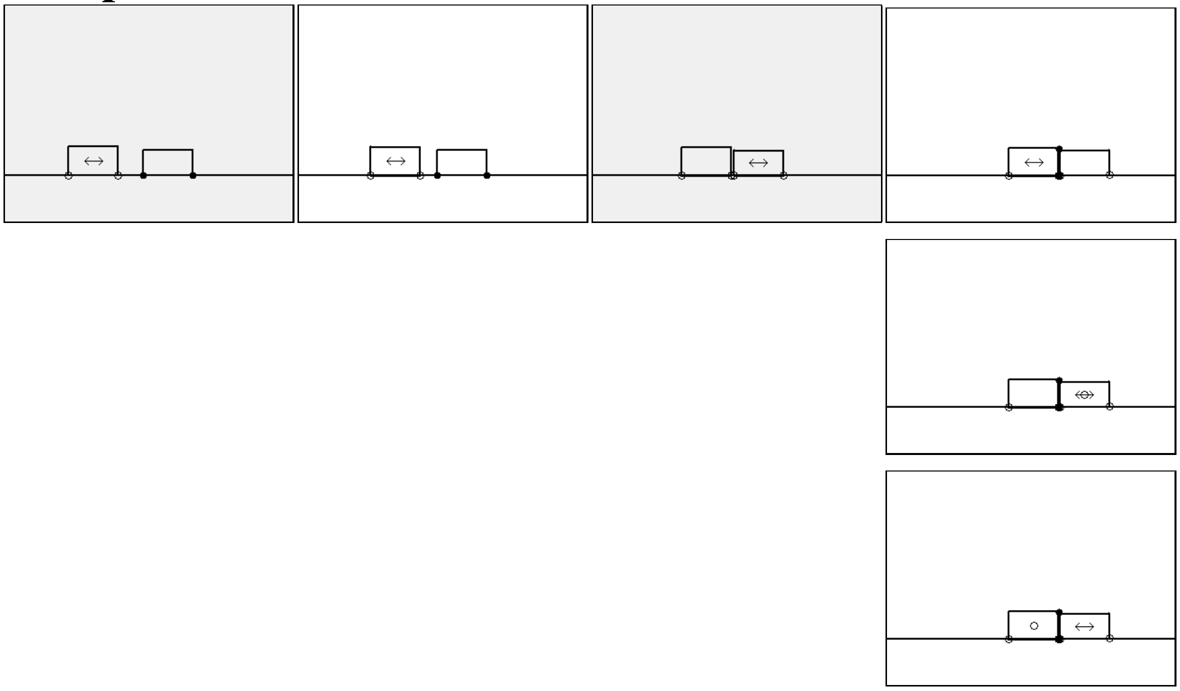
- Track objects (cf. Mann, Jepson, & Siskind, 1996)
- Detect and remove anomalous frames (collisions, start-stops of motion)
- Force-dynamic analysis of each frame
- Find minimal set(s) of object properties that explain sequence

Cars

20 *24* *28* *36*

Sequence interpretation(s): $[\textsc{Driver}(left)]$

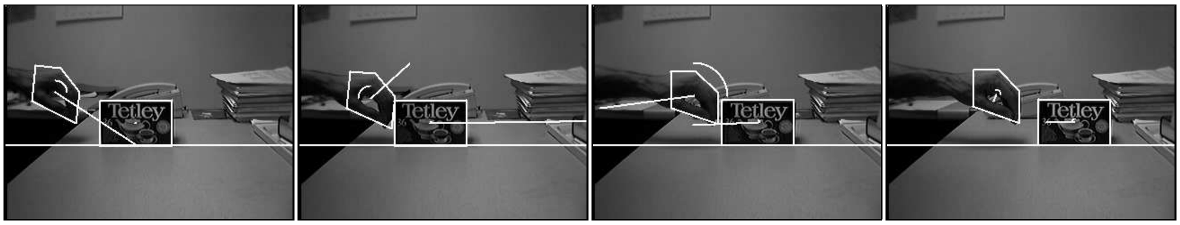
Interpretations of selected frames:



- Frame 28 (the collision) and frame 20 (the start) are *rejected*. (Rejected frames are greyed out.)
- Left-right arrow depicts DRIVER object; small open circle depicts GRASPER object.

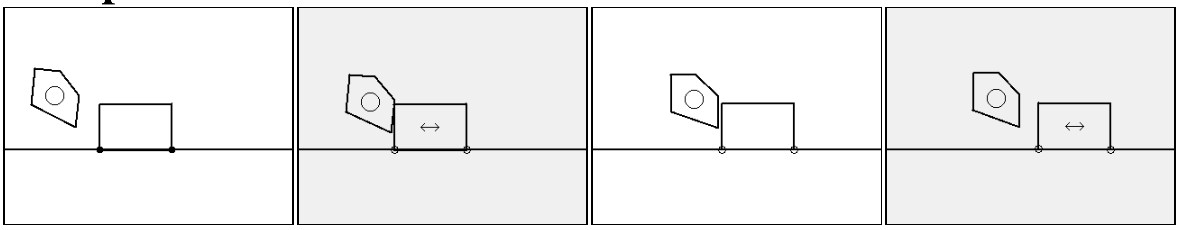
Hit

26 30 34 40

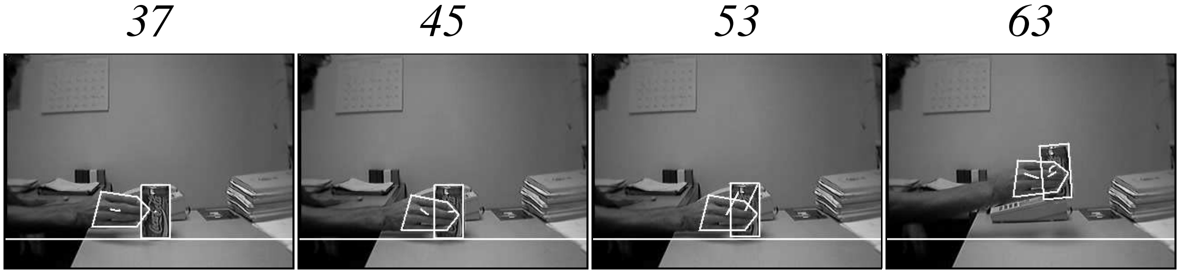


Sequence interpretation(s): $[\textsc{Flyer}(hand)]$.

Interpretations of selected frames:



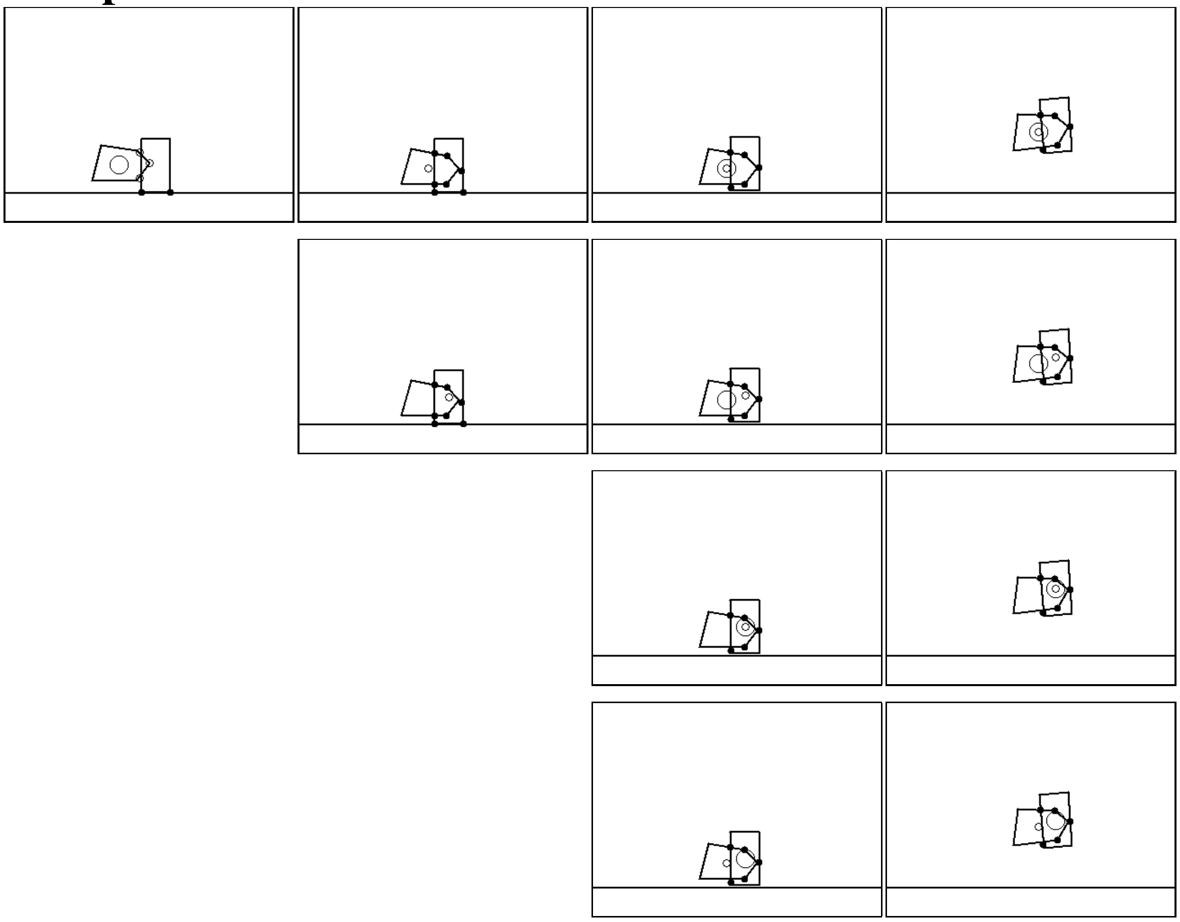
Coke



Sequence interpretation(s):

$[\textsc{Flyer}(hand), \textsc{Grasper}(hand)]$,
$[\textsc{Flyer}(hand), \textsc{Grasper}(can)]$.

Interpretations of selected frames:



Note: System correctly identifies hand as "active" object, but either polygon could be the grasper!

5. CONCLUSION

5.1 Contributions

- Exploit temporal information by integrating inferences from multiple frames.
- Automatic detection and rejection of frames where there are motion discontinuities and/or motion ambiguity.

5.2 Future Work

- Instead of discarding frames at collisions, we should exploit this information. In particular, we require a theory of transfer of *momentum* at collisions.
- Simplistic processing assumes properties *persist* for entire sequence.
  (Don't allow for changing properties, appearance and disappearance of objects, merging and splitting, etc.)
- Furthermore, we assume object properties are *independent*.
  To describe *action* in general, a richer set of descriptions is probably needed. (For example, objects with internal state, goals, intentions, etc.)