Towards the Computational Perception of Action*

Richard Mann* Allan Jepson†

*NEC Research Institute, Inc., 4 Independence Way, Princeton, NJ 08540 USA

†Department of Computer Science, University of Toronto, Toronto M5S-1A4 CANADA

Abstract

Understanding observations of interacting objects requires one to reason about qualitative scene dynamics. For example, on observing a hand lifting a can, we may infer that an 'active' hand is applying an upwards force (by grasping) to lift a 'passive' can. In previous work [6] we presented a system that infers qualitative scene dynamics from the instantaneous motion of objects. However, since that analysis only considered single frames in isolation, there were often multiple interpretations for each frame. In this work we show how the dynamic information inferred at each frame can be integrated over time to reduce ambiguity. Our approach to integrating information is to extend our representation to describe objects by a set of properties *or* capabilities *that are assumed to persist over time. Given this extended representation we find interpretations that require the smallest set(s) of properties over the whole image sequence.*

1 Introduction

We consider the perception of simple actions in video sequences. An example of the type of the problem we are considering is shown in Figure 1. In this sequence, a hand reaches for, grasps, and lifts a coke can off a table top. Given this sequence, our eventual goal is to have a computational system that produces conceptual descriptions of the observed actions, such as "the hand *lifts* the coke can". In order to perform this type of inference the system requires at least three components. First, it must have some representation of the qualitative scene dynamics. That is, the system must be able to infer the basic generation and transfer of forces among the participant objects. Second, the system must have an understanding of how the object behaviors relate to the scene dynamics. That is, it must be able to infer (either by recognition of the objects or by observation of their behaviors over time) that the hand is an 'active object' that can generate forces and grasp other objects, while the can is a 'passive object' that is acted upon by the hand. Finally, in order to build conceptual descriptions from image sequences, we need a way to translate the inferred physical descriptions of actions to natural event categories such as "lift", "push", "drop", etc.

*This work was performed while the first author was at the University of Toronto.

Most previous work on motion understanding has focused on only the third problem, recognizing events from image sequences. Unfortunately, these approaches have not used dynamic information. Instead, they have either relied on specific domain knowledge (such as traffic scenes [7]), or have used some form of recognition model based on either predefined templates [4] or hidden markov models [9, 2].

In constrast, we focus on the first two components above. In particular, we attempt to perform a bottom-up inference of physical descriptions of the actions depicted in image sequences in terms of the *force-dynamic* properties of objects.

2 Integrating information over time

In earlier work [6] we presented a system that made inferences of scene dynamics based on the instantaneous motion taken at particular frames of an image sequence. Our approach is based on the analysis of the Newtonian mechanics of a simplified scene model. In particular, objects are modeled as 2D polygons in a layered depth model. Objects are subject to forces due to gravity and contact with other objects. Objects can also attach to other objects in various ways. Finally, in addition to passive behaviors, there can be extra force and torque generators called 'motors' that act either on objects, or at the contact regions between objects. At each frame of the input sequence, our system uses a set of preference rules to find the smallest set(s) of motors and attachments that could produce the observed motion but still be consistent with Newtonian mechanics.

While promising, the previous approach suffers from two fundamental limitations. First, because we consider single frames in isolation, when objects interact there is often ambiguity as to which object (if any) is responsible for generating the forces ('motors') we infer in the scene. For example, at a single frame during the *lifting* phase of the coke sequence (frames 53–63) we cannot determine which object is responsible for generating the forces. Second, since our analysis is based on the instantaneous motion estimated at each frame, anomalous interpretations will occur when there are motion discontinuities (*eg* at collisions or changes in contact geometry) or sign changes in velocity (*eg* at starts and stops of motion).

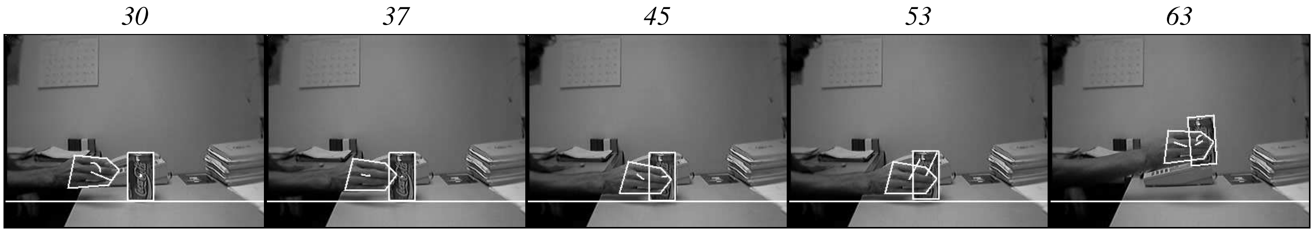


Figure 1: coke sequence: A hand reaches for, grasps, and lifts a coke can off a table top. Polygonal outlines of the objects and estimates for their linear and angular accelerations are overlayed on each frame. (The horizontal line denotes the table top.)

2.1 Ascribing properties to objects

Our approach to integrating information is to describe objects by a set of *properties* or *abilities* that are assumed to persist over time. For example, given successive observations of a hand moving freely or imparting forces on other objects in the scene, we would like to infer that the hand has the ability to move autonomously. Similarly, given successive observations of a hand which is lifting or pulling other objects, we may infer that the hand has an ability to grasp other objects.

Specifically, our representation includes the following *object properties*:

- FLYER(o) — object o can generate an arbitrary force and torque on itself;
- DRIVER(o) — object o can exert an arbitrary tangential force along any of its edges;
- ROTOR(o) — object o can exert an arbitrary torque at any of its vertices;
- GRASPER(o) — object o can attach to any object which contacts it in the scene.

where o refers to any object in the scene. Given this representation, we seek interpretations that require the smallest set(s) of properties *over the whole sequence*. As in [6], we assign *priorities* to minimization of the various properties. Specifically, we minimize flyer properties at the highest priority, followed by driver and rotor properties at the next priority, followed by grasper properties at the lowest priority.

2.2 Discontinuity detection

As described above, our system assumes that objects are in continuous motion, and does not model collisions or force impulses. Therefore, in addition to integrating information over time, we need a way to select only those frames where the motion is continuous.

In this work we assume that all force impulses (step changes in velocities) are due to collisions. Furthermore, we assume that all such collisions occur among visible (non-occluded) objects. To detect collisions, we implement the simple test shown in Figure 2. Specifically, our system looks for any impending *penetration* (negative relative velocity normal to the contact region) between two abutting objects. As shown in Figure 2, the collisions can be of two sorts. The first collision type (a "push") occurs when the objects remain in contact after the collision. The second collision type (a "bounce") occurs when the objects separate immediately after collision. Note that a discontinuity is *not* detected when two objects in different depth layers move smoothly past each other. [1]

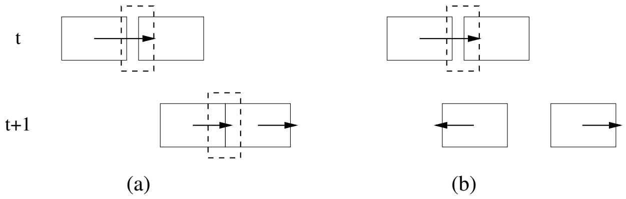


Figure 2: Collision detection process. Collisions are detected when an impending penetration is followed by either an abutting contact (a), or a separation (b). The arrows depict the object velocities. The dashed box depicts the contact tolerance between the objects.

The tests are implemented by comparing contact regions in adjacent frames of the sequence. Upon detection of a collision, the two adjacent frames are removed. In addition, because the motion estimates are unreliable in the neighborhood of a collision, we removed one preceding and two subsequent frames.

In addition to removing discontinuities due to collisions, we also need to remove frames at which there is a

[1] Note that the contact test must also consider cases where the contact is maintained, but changes geometry. This occurs, for example, when an object tips towards or away from a surface (see Figure 7). The full details of the implementation are described in [5].

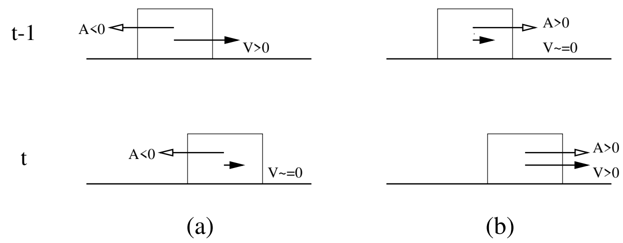


Figure 3: Sliding objects. (a) depicts an object that is decelerating to rest (stopping), while (b) depicts an object that is accelerating from rest (starting). Stopping can usually be explained by sliding friction (a tangential force opposing the direction of motion), while starting requires the presence of a motor to generate a forward force. When the velocity becomes small, so that the actual sign is uncertain, then we cannot distinguish between cases (a) and (b) given data at a single frame.

significant tangential acceleration, but the sign of the tangential velocity cannot be resolved. The reason is that, when dealing with just the instantaneous motion, the relative signs of the tangential velocity and acceleration dictate whether the motion corresponds to a stop or a start (see Figure 3). In situations where the velocity is too small to be able to reliably infer the sign, we therefore can easily confuse stops and starts. Confusing a stop with a start may cause our system to infer the anomalous presence of a motor. To avoid this, we simply remove all frames containing such starts and stops.

Once the offending frames are removed from the sequence, we find minimal sets of object properties sufficient to explain the motion in each frame. Intuitively, we can view the temporal integration process as an incremental search that starts with the assumption that none of the objects have any special properties, and gradually increases the set of object properties as more behaviors are observed. Note that, unlike the earlier system that performed the search on single frames, for each set of properties chosen, all valid frames of the sequence are tested.

3 Results

We tested the system given tracking data from several sequences (*cf* [6]). Figure 4 shows the results for the coke sequence. Below each frame of the sequence we show the minimal set(s) of object properties (FLYERs, DRIVERs, ROTORs, and GRASPERs) sufficient to explain the instantaneous motion in that frame. At the bottom of the figure we show the minimal set(s) of object properties sufficient to explain the *entire sequence*. Note that while the system has inferred that the hand is a flyer, there are multiple interpretations since it cannot determine whether the hand or the can is a grasper object.

Figure 5 shows the results for the cars sequence. In this sequence a hand releases a wind up car that accelerates, hits, and then pushes a second car. In this case, while there is still ambiguity at some individual frames (*eg* frame 36), temporal integration results in a single interpretation in which the left car is a driver object. [2] As described in §2.2, our system removed the frames around frame 20 (start of the left car) and frame 28 (the collision) before integrating the results over time. Note that removal of the collision was necessary to avoid anomalous interpretations for this sequence.

Figure 6 shows results for the hit sequence. Here a hand hits a box that is sitting on the table top. The box slides and then comes to a stop. Again, a unique interpretation is found since the motion can be explained by an active hand. As expected, the deceleration of the box is explained by sliding friction on the table top. Removal of the frames around both the collision (frame 30) and the stop (frame 40) were necessary to obtain a unique interpretation for this sequence.

Finally, figure 7 shows the results for the tip sequence. In this sequence a hand raises a box onto its corner and allows it to tip to an upright position. The system removed frames around frame 30 (hand starting to slide against the table) and frame 38 (collision between the box and table). In this case the temporal integration process is only partially correct: While the dynamics analysis has recognized the passive motion of the tipping box, it has not been able to detect the 'active' hand. The problem is that while the displacement of the hand near frame 30 changes over time, the velocity is too small to reliably detect the direction of motion, so this frame is removed.

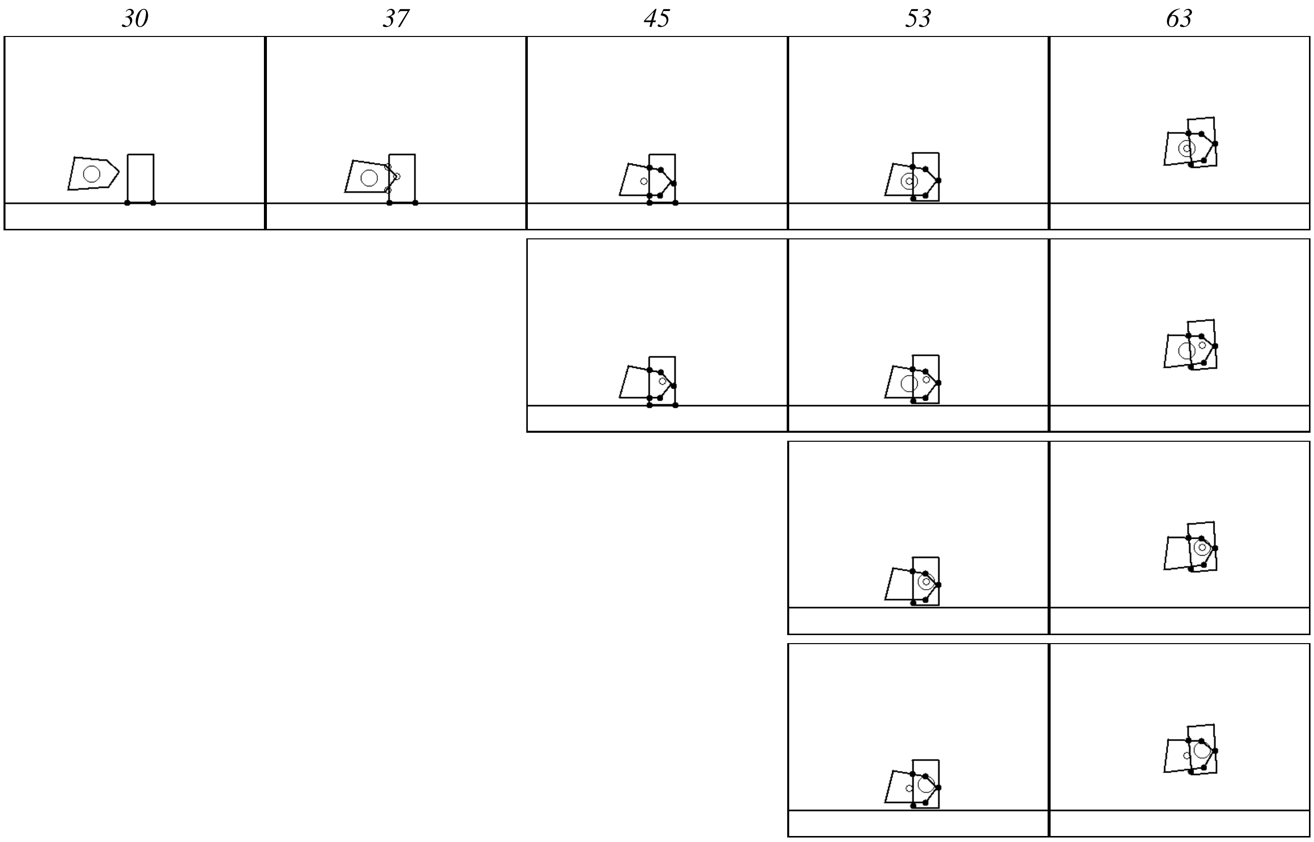
4 Discussion and Conclusions

We have shown that to extend our previous system [6] to integrate information over time we need to: 1. Remove frames where the instantaneous motion is either discontinuous, or contains insufficient information to reliably identify the object properties. 2. Ascribe properties to objects and minimize these properties over the entire sequence.

Our approach to continuous sequence processing by choosing minimal sets of properties over time is similar to [8], except that instead of using qualitative physics, our system uses an explicit theory based Newtonian mechanics. Our approach to discontinuity detection is similar to that described in [1] except rather than detecting and classifying isolated frames containing discontinuities, we attempt to infer object properties over the entire sequence.

There are a number of natural extensions to this work.

[2] Note that we have disallowed DRIVERs (or any other properties) for the table top.



Sequence Interpretation(s): [FLYER$(hand)$, GRASPER$(hand)$], [FLYER$(hand)$, GRASPER(can)].

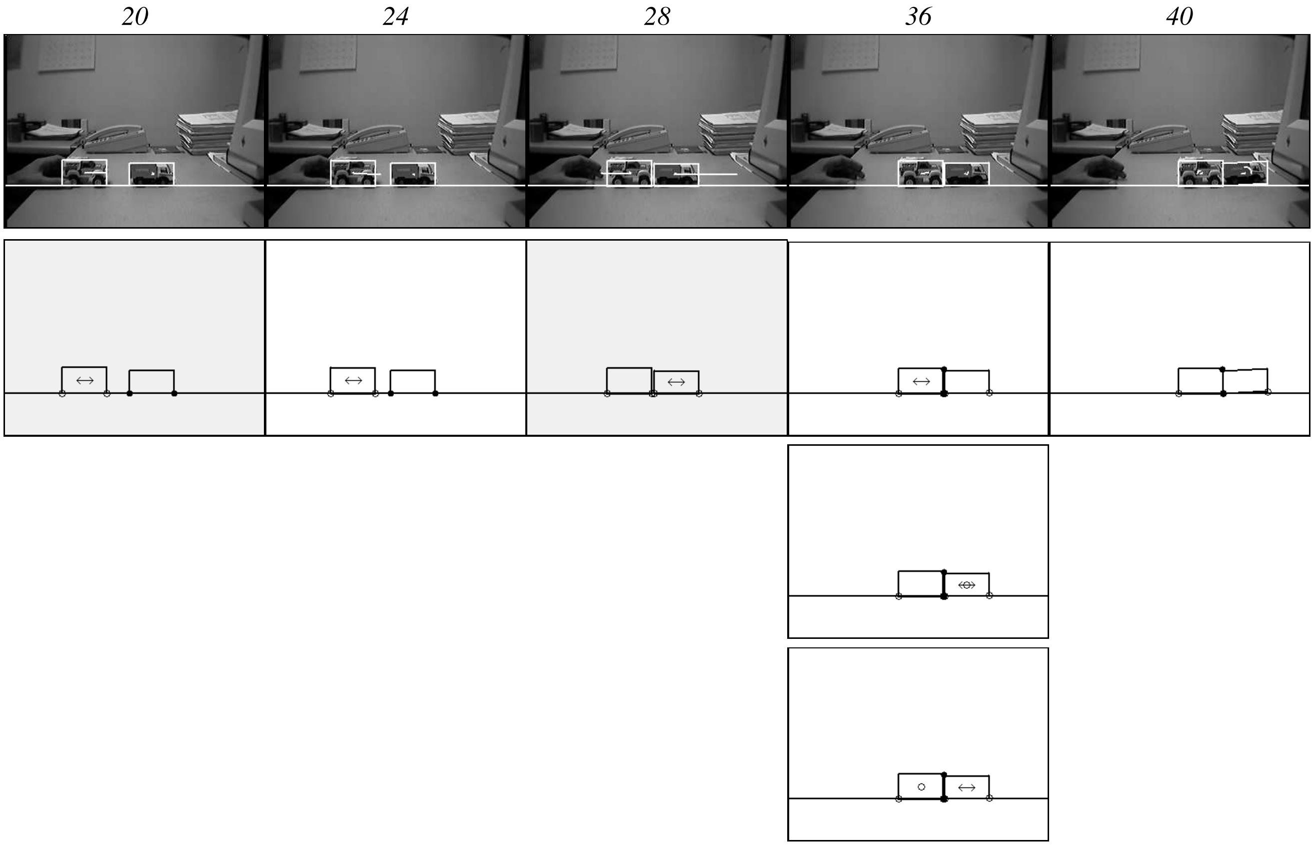
Figure 4: Object properties and temporal integration results for the sequence in Figure 1. Minimal set(s) of *object properties* are shown for each frame of the sequence. A large open circle at the object center denotes a FLYER object while a small open circle denotes a GRASPER object. Minimal set(s) of object properties that explain the entire sequence are shown at the bottom of the figure. Note that since neither object is moving in frame 45, no FLYER assertions are required in that frame.

First, rather than throwing away frames where the instantaneous dynamic model breaks down, we should attempt to exploit this information. In particular, it should be possible to extend our sequence processing by looking at frames in the neighborhood of motion discontinuities to determine the category from amongst several possibilities (collisions, starts and stops). Such a set of motion categories is presented in [3], however, these categories would have to be expanded to deal with polygons contacting in various ways. To handle these cases would also require extending our dynamic model to deal with collisions by looking at the velocities immediately before and after a collision and analyzing the transfer of momentum between the objects.

Second, our representation of object properties is still simplistic. In particular, in our representation, the object properties are taken to be *independent*. Because of this, there will be multiple interpretations whenever objects share attachment or motor assertions (*eg* the coke sequence). In addition, because our system considers the minimal set of properties for the entire sequence, there is no provision for object properties that change with time, or for the creation, deletion, or merging of objects in the scene. To deal with these cases will likely require either additional structure within the representations, or the observation of objects over a wider range of behaviors. In particular, such structure should allow the representation of various internal states, intentions, and goals for the participant objects.

Finally, as described in the Introduction, we need a way to translate the inferred force-dynamic descriptions into natural event categories. (See [8] for such a proposal based on recognizing specific sequences of force-dynamic descriptions.)

In summary, while the results reported here are preliminary, they indicate that the integration of information over time can significantly reduce ambiguity in sequence interpretation. However, much work remains to be done to de-



Sequence Interpretation(s): [DRIVER($left$)].

Figure 5: Object properties and temporal integration results for the cars sequence. Minimal set(s) of *object properties* are shown for each frame of the sequence. A left–right arrow is used to denote a DRIVER object while a small open circle is used to denote a GRASPER object. The minimal set(s) of object properties that explain the entire sequence are shown at the bottom of the figure. Frame 20 (start of the left car) and frame 28 (the collision) have been removed by the temporal integration procedure (removed frames are shown in grey). Note that the interpretation at frame 28 is *anomalous* because of unreliable estimates for the accelerations at the collision.

velop this into a complete event recognition system.

References

[1] Matthew Brand. Understanding manipulation in video. In *Proceedings, 2nd International Conference on Face and Gesture Recognition*, Killington, VT, 1996.

[2] Matthew Brand and Nuria Oliver. Coupled hidden markov models for complex action recognition. In *CVPR96*, 1996.

[3] Allan D. Jepson and Jacob Feldman. A biased view of perceivers. In David Knill and Whitman Richards, editors, *Perception as Bayesian Inference*, pages 229–235. Cambridge University Press, 1996.

[4] Yasuo Kuniyoshi and Hirochika Inoue. Qualitative recognition of ongoing human action sequences. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 1600–1609, Chambery, France, August 1993.

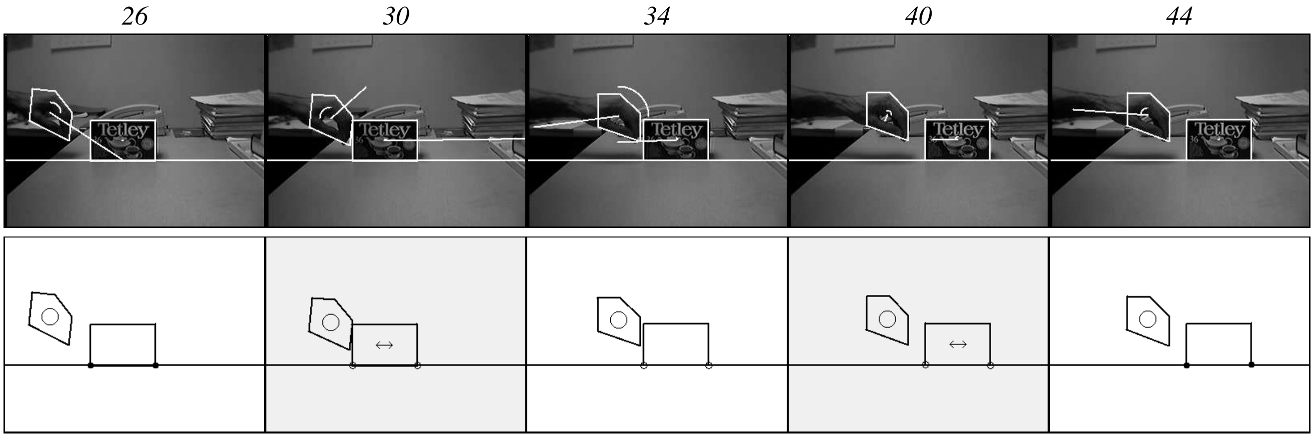
[5] Richard Mann. *Computational Perception of Scene Dynamics*. PhD thesis, Department of Computer Science, University of Toronto. In preparation.

[6] Richard Mann, Allan Jepson, and Jeffrey Mark Siskind. The computational perception of scene dynamics. *Computer Vision and Image Understanding*, 65(2), February 1997.

[7] H-H Nagel. From image sequences to conceptual descriptions. *Image and Vision Computing*, 6(2):59–79, May 1988.

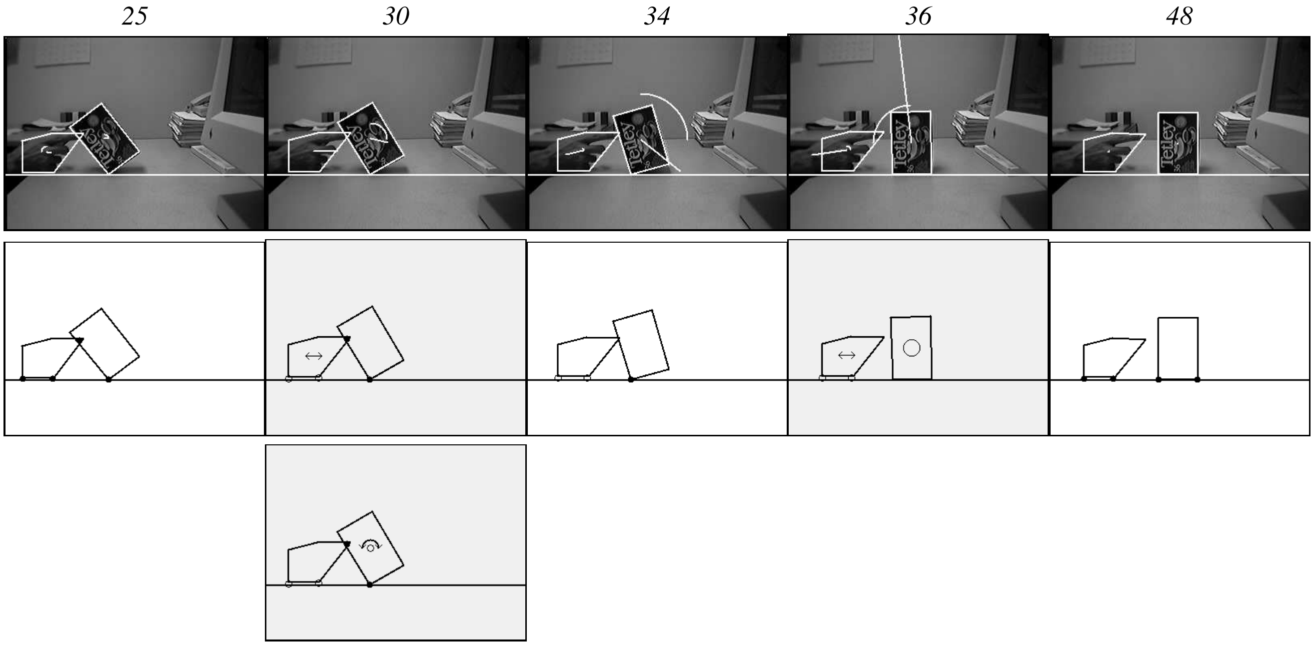
[8] Jeffrey Mark Siskind. *Naive Physics, Event Perception, Lexical Semantics, and Language Acquisition*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, January 1992.

[9] Jeffrey Mark Siskind and Quaid Morris. A maximum-likelihood approach to visual event classification. In *ECCV96*, pages II:347–360, April 1996.



Sequence Interpretation(s): [FLYER($hand$)].

Figure 6: Temporal integration results for the hit sequence. A single interpretation is found: all motion can be explained by an active hand. Frames 30 (a collision) and 40 (a stop) were removed by the temporal integration procedure. Note that the interpretation at frame 30 is *anomalous* because of unreliable estimates for the accelerations at the collision.



Sequence Interpretation(s): [].

Figure 7: Temporal integration results for the tip sequence. In this case a single (incorrect) interpretation is found for the sequence: all objects are passive. See text for details. Frame 30 was removed due to the start of motion between the hand and the table. Frame 36 was removed due to the collision between the box and the table. The curved arrow at the object center denotes a ROTOR object. Note that only two of the five possible interpretations for frame 30 are shown here. (The three additional interpretations are obtained by allowing either the hand or the box to be a GRASPER object and by allowing the hand to be a ROTOR object.)