The Computational Perception of Scene Dynamics*

Richard Mann, Allan Jepson[†] and Jeffrey Mark Siskind
Department of Computer Science
University of Toronto
Toronto Ontario M5S 1A4 CANADA
Phone: 416/978–6114, FAX: 416/978–1455
Email: {mann, jepson, qobi}@cs.toronto.edu

[†]also at Canadian Institute for Advanced Research.

November 2, 1995

Abstract

Understanding observations of interacting objects requires one to reason about the force-dynamic relations between objects. We present an implemented computational theory that derives force-dynamic interpretations directly from camera input. Interpretations are expressed in terms of assertions about the kinematic and dynamic properties of objects. The feasibility of interpretations can be determined relative to Newtonian mechanics by a reduction to linear programming. Multiple feasible solutions are compared using a preference hierarchy to select plausible interpretations. We provide computational examples to demonstrate that our ontology is sufficiently rich to describe a wide variety of image sequences.

KEYWORDS: Motion understanding, Scene dynamics, Perceptual inference, Knowledge-based perception, Domain theory, View-based representations.

*Submitted.

1 Introduction

Both AI and psychology researchers have argued for the need to represent "causal" information about the world in order to make inferences. In particular, understanding motion sequences requires the observer to postulate forces on objects and force transfer between interacting objects. In this paper we make these ideas precise with an implemented system that can make causal inferences directly from video sequences.

The use of domain knowledge by a vision system has been studied extensively for both static and motion domains. Our work addresses a number of important limitations in prior work:

- Our system embodies a sound inference procedure based on an explicit physical theory.
- Our system models kinematic and dynamic information about the world in addition to static information available from a single frame.
- Our system can represent uncertainty to make inferences from ambiguous input.
- Finally, our system performs such inferences directly from camera input.

We postpone a more detailed discussion of related work until §8.

In particular, this paper makes three central contributions. First, we provide an ontology suitable for producing interpretations of image sequences in terms of the kinematic and dynamic properties of observed objects. Second, we provide a computational procedure to test the feasibility of such interpretations by reducing the problem to a feasibility test in linear programming. Finally, we provide a theory of preference ordering between multiple interpretations along with an efficient computational procedure to determine such orderings.

2 Overview

It is useful to first present a brief overview of the main components of our system. Here we emphasise the basic intuition behind each of the components and the prerequisites for their use. In later sections we describe each of the components in detail, and present experimental results.

We provide this overview in terms of a single illustrative example, the `coke` sequence, shown on the top line of Figure 1. (The set of sequences displayed there are used for the computational examples in §7.) In particular, a hand is reaching for, grasping, and then lifting a coke can off of a desk top. As mentioned in the introduction, our eventual goal is to have a machine vision system that, when given image sequences such as this one, can understand the basic force generation and force transfer relationships of the various objects in the scene. For this example sequence, the system should understand that the can is initially supported by the table and that the hand (and arm) is possibly an 'active object.' Roughly speaking, an active object is something that can generate forces other than those due to gravity, friction, and acceleration. Objects that are not active, such as the table and

Figure 1: The example sequences: coke, cars, hit, arch, and tip. The frame numbers are given above each image.

the coke can, are said to be 'passive objects.' Later, during the lifting phase of this sequence, if the system remembers that it has determined that the hand is an active object, then it should conclude that the hand is attached to the coke can (i.e. grasping it) and applying an upward force on it.

This goal of understanding forces and dynamics given image sequences is an ambitious one. Obviously we do not achieve it completely here. However, we do present an implementation that takes us much of the way to our goal, albeit in a simple domain. The critical simplification is that the objects and their motions can be approximated by a layered 2D representation. Such a representation provides a crude but useful approximation for a variety of situations, including all of the sequences depicted in Figure 1. Moreover, the simplicity of the 2D domain serves to highlight the basic principles behind our approach. It is our belief that these principles can also be applied to general 3D problems.

2.1 An Outline of the Approach

In order to reason about dynamic properties of the objects in the scene we use the equations of Newtonian mechanics. Given a concrete hypothesis about the scene content, we simply express the corresponding equations of Newtonian mechanics in a suitable form and check to see if they have a feasible solution.

This physics-based approach raises stringent requirements on the richness and the level of detail of the underlying scene model. Indeed, we have the basic requirement that the geometry of the scene must be completely specified in order to express these equations. That is, we need to know object shape and placement, along with the surfaces of contact between objects. In addition, we need estimates for the velocities and accelerations of the observed objects. We also need to assume something about the distribution of mass within each object. For our purposes here, we simply take an object's center of mass to be it's centroid, as if it had a uniform mass distribution. The total mass and the inertial matrix are then treated as constrained unknowns when the system checks for a feasible solution. In summary, a strong prerequisite on the hypothesis-generation mechanism is therefore that a rather detailed description of the scene is needed before the physics-based modeling can be applied.

It is convenient to define a *configuration* to be the set of scene properties that are necessarily present, given the image data and any restrictions inherent in the ontology. For example, in the current system, the positions, velocities, and accelerations of the objects are provided by the image observations, and the positions of the centers of mass are fixed, by our ontology, to be at the object centroids.

For a concrete example, consider the lifting phase in the coke sequence. Through the use of a tracking algorithm, we obtain the 2D positions, velocities, and accelerations of polygons that roughly describe the shape of the scene objects. This information, along with the line denoting the table surface, constitutes the configuration. The polygons provided by the tracker are displayed in Figure 2, along with other symbols, discussed further below, that denote additional properties of the various interpretations.

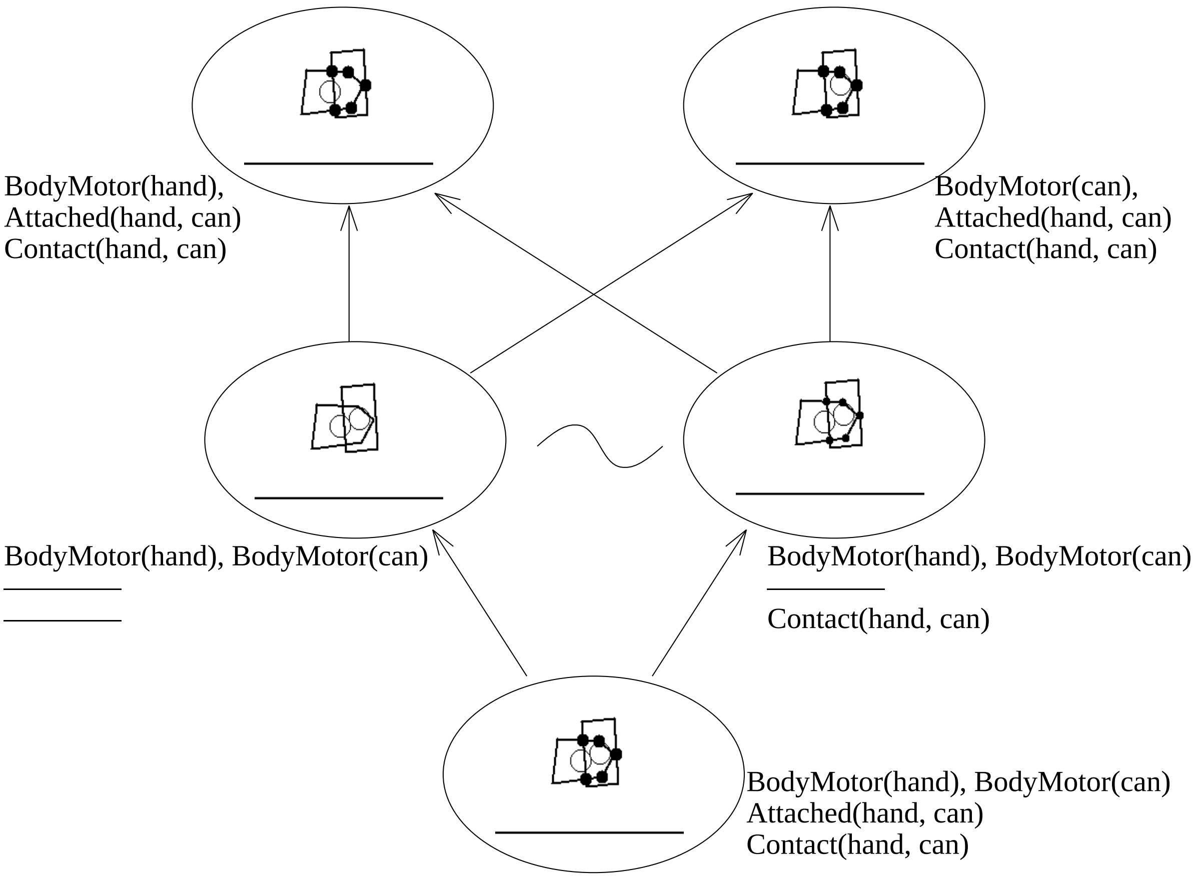

Figure 2: The preference ordering for the five feasible interpretations of the lifting phase is displayed. A large open circle at object center denotes a BodyMotor, and the large, medium, or small black disks at vertices denote Attach, Contacting, or not Contacting objects, respectively. A textual form of the assertions appears adjacent to each interpretation. The three levels of priority are represented by each line of text. Note that the absence of an assertion denotes it's negation.

In order to supply the information missing from a configuration, we consider *assertions* taken from a limited set of possibilities. These assertions correspond to our hypotheses about the basic force generation and force transfer relationships between objects. For brevity, we constrain ourselves here to only those assertions that play a significant role in our coke sequence. The full set is provided in §3. In particular, here we need the following three types of assertions:

- $\textsc{Contact}(o_1, o_2, c)$ — objects o_1 and o_2 contact in the scene with the region of contact c;
- $\textsc{Attach}(o_1, o_2, P)$ — objects o_1 and o_2 are attached at some set P of points in the contact region;
- $\textsc{BodyMotor}(o)$ — object o has a 'body motor.'

The intuitive meanings of these assertions are: (i) objects o_1 and o_2 are contacting, either in depth or abutting; (ii) objects are attached on some set P of points in the contact region; these attachment points are functionally equivalent to rivets, fastening the objects together; (iii) the object can generate an arbitrary force and torque on itself, as if it had several thrusters. Note that the attachment assertion is properly understood to be a characterization of what sorts of forces are supported at the attachment points. Attached objects can be pulled, pushed, and sheared without coming apart while, without the attachment, the contacting objects may separate or slide on each other depending on the applied forces and on the coefficient of friction. The precise choice of these kinematic and dynamic assertions is not critical for the purpose of this paper. What is important is that they are sufficiently rich to describe a wide variety of phenomena.

Note that, given the hand and can polygons in Figure 2, there is no evidence that any of these assertions are individually true or false. It is for this reason that they, or their negation, do not belong to the configuration. Furthermore, notice that some sets of assertions are not *admissible*, in that they violate basic constraints for their application. For example, two objects cannot be both attached and not contacting. Moreover, some sets of assertions are not *complete* in that they leave some properties unspecified. In order to check for a feasible force balance, however, our physical model requires the set of assertions to be complete (i.e. it must fully specify each admissible assertion as either true or false). Next, for any complete and admissible set of assertions, we can check the feasibility of those assertions according to whether or not a force balance exists in our physics-based model. This feasibility test can, to a good approximation, be reduced to a linear programming problem (see §5). By definition inadmissible assertion sets are also infeasible. We define an *interpretation* to be a combination of a configuration and a feasible set of assertions which completely specify the scene.

For the lifting phase of the coke sequence there turn out to be five feasible interpretations, within our ontology, as displayed in Figure 2. Note that each of these interpretations can be seen to have a feasible force balance, given that we can adjust both the masses and the forces generated by the body motors. In contrast, any interpretation without a body motor is infeasible since the vertical acceleration cannot be 'explained' (i.e. balanced with a corresponding force). Similarly, the active hand in the top left interpretation in Figure 2 must be attached to the can since, in our ontology, we do not model forces in depth. This means, in turn, that the physics-based model cannot generate a frictional force between the hand and the can. So, without attachment, the observed vertical acceleration of the can cannot be balanced with a vertical force. While this sort of 'reasoning' seems sophisticated, it is simply relying on the feasibility of linear programming problems that the algorithm can pose for itself.

A naive system would generate all feasible interpretations thereby producing all (and just) the interpretations depicted in Figure 2. Our algorithm, however, does not generate all such feasible interpretations since, as we discuss next, many of these interpretations are not of interest. It turns out that the algorithm need not evaluate all interpretations in order to find the interesting ones.

2.2 Preferred Explanations

Up to this point we have discussed the nature of an interpretation in terms of a scene configuration along with assertions that specify further dynamic and kinematic properties. We pointed out that the feasibility of such an interpretation can be checked using a physics-based model for the forces involved along with the admissibility of the assertions. However, given a fairly rich ontology, it is common for there to be multiple feasible interpretations for a given scene configuration (see Figure 2).

Indeed, given that a body motor is capable of generating an arbitrary force and torque on the object possessing it, it follows that for any scene configuration there is always at least one trivial interpretation in which every object has a body motor. The bottom three interpretations in Figure 2 are examples of such trivial solutions. They are guaranteed to pass the force-balance feasibility criterion and, as such, are not very informative interpretations on their own. Rather, we seek interpretations that require, in some specified sense, the *weakest* properties of the various objects.

Model preference relations, as discussed by Richards, Jepson, and Feldman (to appear), can be used to express suitable preference orderings. The basic idea is simple, namely to compare two different interpretations in terms of a prioritised set of elementary preference relations. Our current ontology includes the elementary preference for the absence of the assertion BodyMotor(o) for each object o. In other words, we prefer to see a particular object as a passive object, if feasible, given that other elementary preference relations do not contradict this preference. In addition, we prefer the absence of a kinematic constraint for attachment between two objects over the presence of such a constraint. Finally we are indifferent as to whether we assert that two objects are contacting or not (always assuming admissibility and feasibility).

These three elementary preference relations are taken to be stratified in terms of priority. In particular, the preference against the presence of a body motor is taken to have a higher priority than the elementary preference for no attachment. Finally, we have no elementary preference relation for contact over no contact—these two situations are considered to be equally preferable. Note that this priority ordering is indicated in Figure 2 by placing the assertions adjacent to each interpretation. Assertions are placed on different lines to indicate their priority. It is convenient to place contact assertions on the bottom row even though there is no preference between different contact assertions.

Given these three elementary preference relations along with the priority ordering, the induced ordering on our five feasible interpretations for the coke example is provided simply by a prioritised subset ordering. In particular, suppose one is given two feasible interpretations along with their corresponding assertion sets. If, at the highest level of priority, one interpretation's assertion set is a strict subset of the other's, then that interpretation is preferred. For example, the top left interpretation in Figure 2 is preferred over the three lowest interpretations because it does not assert a body motor on the can. In a sense, not needing a body motor is considered to be major simplification in what the various objects are asserted to be capable of doing, and thus the preference against body motors is placed at a high priority.

Alternatively, if the assertion sets at the highest priority do not satisfy a subset relation then the interpretations are considered to be unordered. For example, the top two interpretations in Figure 2 are considered unordered since the body motor is on different objects. If the sets at the highest priority are the same, then we check the assertions at the next lower priority. For example, the three lower interpretations in Figure 2 all contain the same motor assertions, but can be further ordered based on minimising attachment. Finally, if interpretations agree on both motor and attachment assertions, we are indifferent as to the preference relation. This occurs in Figure 2 for the middle two interpretations, since these differ only in terms of their contact relations.

This approach to preference ordering will not always yield a unique preferred interpretation. Therefore, multiple percepts are possible. For example, in the coke sequence there are two maximally-preferred interpretations, namely either the hand is lifting the can, or vice versa. Given that our ontology includes nothing about hands or cans, that is, we just have moving polygons and notions of mass and force, arriving at these two interpretations for this single frame is intuitively the right thing to do. Indeed, Jepson and Richards (1993) and Richards, Jepson, and Feldman (to appear) propose that such maximal interpretations provide a computational model for a 'percept.' Moreover, Richards, Jepson, and Feldman (to appear) explore the relationship between such preference orderings and qualitative probabilistic models (see also Jepson, Richards, and Knill, to appear).

2.3 Implementation and Limitations

Our current implementation has a number of limitations. First, it is restricted to a 2D layered representation. Thus our system can currently process only fronto-parallel scenes and cannot reason about occlusion or motion in depth. Second, the system does not integrate information over time or reason about object capabilities. In particular, consider the coke sequence once again. During the reaching phase of this sequence, the hand is seen to be above the table and accelerating upwards. Our system concludes that the hand must be an active object while the coke can could be passive and at rest on the table. This is the unique maximally-preferred solution for this frame. As we have just shown, however, there are multiple interpretations of the scene during the subsequent lifting phase; either the hand or the coke can could be active, but not both. Since our system lacks prior knowledge of object capabilities (such as the fact that hands are typically active, and coke cans are typically passive), and does not integrate conclusions over time (such as noticing that it previously concluded the hand must be active), the system is left with multiple interpretations for the lifting phase. Despite these limitations, our system contains a sufficiently rich ontology to describe all of the movies in Figure 1.

We believe that these limitations are not fundamental to our general approach. In particular, our current system is able to reason about 3D representations given suitable 3D input. We are currently investigating approaches for extracting such 3D information from images such as the methods described by Black and Jepson (submitted).

3 Ontology

In this section we discuss the details of the representation that is used in the implementation and motivate some of the choices made. As discussed in §2, our current implementation is limited to a layered 2D representation of the various objects in the scene. This is a reasonable approximation if we assume that objects move in planes roughly fronto-parallel to the camera. In addition, object velocities and accelerations are taken to be continuous functions of time. In particular, we do not consider force impulses that give rise to step discontinuities in the velocities. Despite these limitations, we are left with a surprisingly rich domain, as is indicated by the variety of computational examples displayed in Figure 1.

The ontology can be divided into geometric and dynamic components as described in the next two subsections.

3.1 Geometric Model

The basic primitive for an object part is a rigid two-dimensional convex polygon. A single *object* is a rigid union of convex polygons.

To represent the spatial relationship between objects in the scene we use a *layered* scene model. In our layered model there is no depth ordering. Instead, we represent only whether two objects are in the same layer, in adjacent layers, or in layers separated in depth. Objects can contact either within the same layer or between adjacent layers. The first type of contact, called *abutting contact*, occurs when two objects in the same layer contact at a point or at an edge along their boundary. The second type of contact, called *overlapping contact*, occurs when two objects in adjacent depth layers contact over part of their surfaces and the region of overlap has non-zero area. We denote both types of contact by $\textsc{Contact}(o_1, o_2, c)$ where o_1 and o_2 are the objects and c is a *contact region* between the two objects.

In addition to position information, the relative motion of objects constrains their allowable contact and layer relations. In the case of abutting contacts, contact is admissible only when the relative velocities of the contacting objects is tangential to the contact region (i.e. objects can slide along their contact region, but cannot penetrate or separate). In the case of overlapping contacts, the objects must be in different layers. Finally, the relative depth of objects that do not overlap or abut is left unspecified.

3.2 Dynamic Model

In order to perform force balancing we need each object's center of mass, total mass, and moment of inertia. To determine these properties from image data, we assume that objects have a symmetrical mass distribution. Thus the center of mass of each object is taken to be at that object's geometric center. Object masses, however, are left as free parameters (constrained to be positive) that can be adjusted by the physics-based model in order to find a feasible solution.

We also need estimates of the inertial tensors for objects. For the case of two-dimensional motion considered in this paper the inertial tensor I is a scalar. In order to reflect the uncertainty of the actual mass distribution, we allow a range for I. The upper bound for I is provided by considering an extreme case where all of the mass is placed at the furthest point from the center. The lower bound is provided by an alternate case where all of the mass is distributed uniformly inside a disk inscribed in the object. Together, these provide the constraint

$$\frac{1}{2} M r_{min}^2 \leq I \leq M r_{max}^2 \tag{1}$$

where r_{min} and r_{max} are the minimum and maximum radii of the object.

An object is subject to gravitational and inertial forces, and to forces and torques resulting from contact with other objects. The dynamics of the object under these forces is obtained from the physics-based model described in detail in §5.

Finally, particular objects may be denoted as *ground*. We typically use this for the table top. Forces need not be balanced for objects designated as ground.

3.3 Assertions and Interpretations

We denote an interpretation constructed from a configuration C and an assertion set A as $i = (C, A)$. For our examples, C consists of the object positions, velocities, accelerations, polygonal shapes, and centroids. The assertions A describe additional kinematic and dynamic constraints on the objects.

Currently, our implementation uses the two types of kinematic assertions, $\textsc{Contact}(o_1, o_2, c)$ and $\textsc{Attach}(o_1, o_2, P)$ defined in §2, along with their negations. The admissibility constraints for contact are discussed in §3.1 above. The admissibility constraints for attachment require that, in addition to contact, two objects not exhibit relative motion and that the set P of attachment points is contained within the intersection of the object polygons.

In addition to kinematic assertions, we have the following three types of dynamic assertions, along with their negations:

- BodyMotor(o) — object o has a 'body motor' that can generate an arbitrary force and torque on itself;
- LinearMotor(o_1, o_2, c) — a linear motor exists between the abutting objects o_1 and o_2. This motor can generate an arbitrary tangential shear force across the motor region c. This region must be contained within the contact region between the objects;
- AngularMotor(o_1, o_2, p) — an angular motor exists at the point p that can generate an arbitrary torque about that point. p must be within the contact region between the objects.

The notion of body motor was introduced in §2. Linear motors are used to generate a shear force across an abutment (providing an abstraction for the tread on a caterpillar). Angular motors are used to generate torques at joints.

We apply the following admissibility constraints to sets of motor assertions. Body motors are admissible on all objects. Linear motors are admissible only at contacts where the direction for application of tangential force can be defined. Thus linear motors are admissible only at point-to-edge and edge-to-edge abutments but not at point-to-point abutments or overlapping contacts. Angular motors are admissible only at a single point within the contact region between two objects and the objects must be attached at this point.

4 Generating Hypotheses

In order to demonstrate the applicability of our approach to camera input we have developed a complete implementation for the simplified domain described in §3. We describe the various components of our implementation below.

4.1 Configuration

To acquire the position and orientation of the object polygons for each frame we use a view-based tracking algorithm similar to the optical flow and stereo disparity algorithms described by Jepson and Black (1993) and Jenkin and Jepson (1994). In particular, a template image is provided for each of the objects, along with a polygonal outline describing the object shape within the template. Given an initial guess for the positions of the objects in the first frame, the tracking algorithm then estimates the two-dimensional position and orientation of these initial templates throughout the image sequence by successively matching the templates to each frame. The position of the object polygons is then obtained by mapping the original outlines according to these estimated positions. Finally, the velocity and acceleration of the polygons are obtained using a robust interpolation algorithm on these position results.

As discussed in the previous section, this position, velocity, acceleration, and shape data constitute the configuration component C of any interpretation. In order to hypothesise an interpretation we therefore need to select an admissible and complete set of assertions A.

4.2 Assertions

In order to explore the space of possible interpretations we must first construct a set of admissible assertions. Given that the allowable forces between objects depend on the contact geometry and the relative motion of the objects (Featherstone, 1987), an analysis of the scene kinematics is necessary. Since we do not have exact shape or motion information, however, we need a way to determine which contact relations are possible. In general, the determination of which contact relations are possible between a set of moving objects is a non-linear programming problem. For the purposes of this paper, however, we implement only a partial test in which we consider only pairwise constraints between objects. We return to this issue in §7.

According to our ontology, contacting objects must either abut or overlap. We can classify contact type by examining the region of intersection between objects. If the area of the intersection region is greater than a specified tolerance the contact type must be overlap. If the area of the intersection region is less than that tolerance, we assume the contact type to be abutment. Overlapping contact is always admissible. Abutting contact, however, must satisfy the additional constraint that the relative motion of the two objects be tangential to the boundary between those objects. Abutting contacts that do not satisfy this constraint are inadmissible. Furthermore, abutment is not admissible for configurations of objects in point-to-point contact except when the objects are attached at the contact point.

An attachment relation between two objects is only admissible when those two objects are in contact and the relative velocities and accelerations of the two objects at the attachment points are less than a specified tolerance. Given contact between two objects, our ontology allows attachment at any set of points in the contact region. To reduce the number of hypotheses that we need to consider, we restrict attachment to the vertices of the perimeter of the contact region. Furthermore, in the system described here we consider only those hypotheses where all such vertices are attached.

Given the contact relations, the admissibility of individual dynamic assertions is determined as follows. $\textsc{BodyMotor}(o)$ is always admissible for any object o. $\textsc{LinearMotor}(o_1, o_2, c)$ is admissible for any pair of abutting objects o_1 and o_2. Finally, $\textsc{AngularMotor}(o_1, o_2, p)$ is also admissible at any pair of attached objects o_1 and o_2, where p is the attachment point. Our current implementation allows angular motors only when the contact region is a single point, that is, when the contact type is either point-to-point or point-to-edge. In this case, the point p is uniquely determined.

Given this specification, it is now possible to generate all admissible interpretations for a given frame of the image sequence. These are obtained from the configuration provided by the tracker combined with every admissible and complete set of assertions.

Among all possible interpretations we must select those that are feasible, namely those that satisfy our physical theory. Nominally we would have to consider preference relations between all feasible interpretations. In practice, however, we avoid generating this full set of feasible interpretations by making use of a monotonicity property of our ontology to reduce the search space. Before we discuss this important property, we first describe the mechanism for checking the feasibility of an interpretation.

5 Feasible Interpretations

Given a configuration of the scene objects along with a set of assertions about the kinematic and dynamic properties of the scene, we can use a theory of dynamics to determine if the interpretation has a feasible force balance. In particular, we show how the test for consistency within the physical theory can be expressed as a set of algebraic constraints that, when provided with an admissible interpretation, can be tested with linear programming. We use a force-balancing approach similar to that proposed by Blum, Griffith, and Neumann (1970). Our approach, however, models dynamics as well as static force balancing.

In this section we present a theory for the general three-dimensional case. The experimental results we describe later were produced using a two-dimensional variant of this theory.

For rigid bodies under continuous motion, the dynamics are described by the Newton-Euler equations of motion (Goldstein, 1980). For rigid bodies of non-varying mass, the appropriate equations are:

$$\begin{aligned} \mathbf{F} &= \dot{\mathbf{p}} \\ &= M\dot{\mathbf{v}} && (2) \\ \mathbf{N} &= \dot{\mathbf{L}} \\ &= \mathbf{I}\dot{\omega} + \omega \times \mathbf{I}\omega && (3) \end{aligned}$$

The first equation relates the total applied force $\mathbf{F}$ to the rate of change of linear momentum $\dot{\mathbf{p}}$. For bodies with non-varying mass, this reduces to $M\dot{\mathbf{v}}$, where M is the object mass and $\dot{\mathbf{v}}$ is the acceleration. The second equation relates the total applied torque $\mathbf{N}$ to the rate of change of angular momentum $\dot{\mathbf{L}}$. For rigid bodies, the rate of change of angular momentum is given by $\mathbf{I}\dot{\omega} + \omega \times \mathbf{I}\omega$, where $\mathbf{I}$ is the inertial tensor, ω is the angular velocity, and $\dot{\omega}$ is the angular acceleration.

Given a scene with convex polygonal objects, we can represent the forces between contacting objects by a set of forces acting on the vertices of the convex hull of their contact region (Featherstone, 1987). Under this simplification, the equations of motion for each object can be written as

$$\begin{aligned} M\dot{\mathbf{v}} &= M\mathbf{g} + \mathbf{F}_b + \sum_{c\in\Gamma} \delta_c \sum_{p\in c} \mathbf{F}_p && (4) \\ \mathbf{I}\dot{\omega} + \omega \times \mathbf{I}\omega &= \mathbf{N}_b + \sum_{c\in\Gamma} \delta_c \sum_{p\in c} \mathbf{N}_p + (\mathbf{p} - \mathbf{r}) \times \mathbf{F}_p && (5) \end{aligned}$$

where $\mathbf{g}$ is the acceleration due to gravity, $\mathbf{r}$ is the center of gravity of the object, and $\mathbf{F}_b$ and $\mathbf{N}_b$ are unknown body forces and torques that may act upon the object. We use the body forces and torques to implement the BODYMOTOR assertion in our ontology. Γ denotes the set of contact regions involving the object. For each contact region $c \in \Gamma$ we add the forces $\mathbf{F}_p$ at each contact point $p \in c$. In addition, we allow terms $\mathbf{N}_p$ at each contact point. These correspond to the ANGULARMOTOR assertions in our ontology. Finally, the terms $\delta_c \in \{-1, 1\}$ encode the direction of the contact forces. The signs of δ_c are arbitrary as long as they are consistent between contacting objects.

5.1 Contact Conditions

Assuming that there are no degenerate contacts we can represent each contact region c by a set of one or more contact points and a vector $\mathbf{n}_c$ normal to the contact region. Contacts that are not asserted to be ATTACHed must obey the normal force constraint

$$\mathbf{F}_p \cdot \mathbf{n}_c \geq 0 \tag{6}$$

where $\mathbf{F}_p$ is the contact force at each point $p \in c$.

In addition, contact points that are not part of a LINEARMOTOR will have tangential forces limited by friction. When the relative velocity at the contact point is zero we constrain the tangential forces by a *Coloumbic friction* model

$$||\mathbf{F}_t|| \leq \mu_r \mathbf{F}_n \tag{7}$$

where $\mathbf{F}_t$ is the tangential component, and $\mathbf{F}_n$ is normal component, of the contact force. The coefficient μ_r is the *coefficient of resting friction*. In the case of objects with relative motion we limit the tangential forces as above, but use a different coefficient μ_s, the *coefficient of sliding friction*. In general we will have $\mu_s < \mu_r$. In addition to the magnitude constraint, we also need to limit the direction of sliding friction to be *opposing* the direction of motion. This is achieved by the additional constraint

$$\mathbf{F}_p \cdot \mathbf{v} \leq 0 \tag{8}$$

where $\mathbf{v}$ is the relative velocity of the contact point projected into the contact plane.

5.2 Testing Feasibility

A configuration is feasible if the motion equations can be satisfied subject to the contact conditions and the bounds on the mass and inertia described in §3.2.

We can approximate these constraints by a set of linear equalities and inequalities as follows. We write the normal force and sliding friction constraints directly. The motion equations can be written as a set of linear equalities by taking the x, y, and z components separately. We can approximate the bounds on the magnitude of the tangential component of the frictional forces by restricting the magnitude in a set of two or more component directions.

Finally, we add conditions on the body forces and torques based on the assertions in A. For passive objects we constrain each of the component directions of the body forces and torques to be zero, within tolerances F_0 and N_0, respectively. Note that the use of tolerances is necessary since the observed motion will never be exactly zero.

Since all of the above equations and inequalities are linear, dynamic feasibility can be reduced to a feasibility test using linear programming (Luenberger, 1984).

6 Preferences

As described in §3, we have a fixed set of elementary preference relations, namely

- $P_{bodymotor}(o) : \neg\text{BodyMotor}(o) \succ \text{BodyMotor}(o)$;
- $P_{linearmotor}(c) : \neg\text{LinearMotor}(o_1, o_2, c) \succ \text{LinearMotor}(o_1, o_2, c)$;
- $P_{angularmotor}(c) : \neg\text{AngularMotor}(o_1, o_2, p) \succ \text{AngularMotor}(o_1, o_2, p)$.

Here $\neg$ denotes the negation of the predicate that follows. These elementary preference relations all encode the specification that it is preferable not to resort to the use of a motor, all else being equal. These elementary preference relations appear at the highest priority (recall Figure 2).

At the next level of priority we have

- $P_{attach}(o_1, o_2, P) : \neg\text{Attach}(o_1, o_2, P) \succ \text{Attach}(o_1, o_2, P)$,

so the absence of an attachment assertion is also preferred.

Finally, at the lowest level of priority, we have the indifference relation

- $P_{contact}(o_1, o_2, c) : \neg\text{Contact}(o_1, o_2, c) \sim \text{Contact}(o_1, o_2, c)$,

so the system is indifferent to the presence or absence of contact, all else being equal.

We wish to find interpretations that are maximally preferred subject to these preference relations. In this paper we consider a special case of the more general orderings described by Richards, Jepson, and Feldman (to appear) where the elementary preference relations can be of the form $P(x) \succ Q(x)$ for predicates P and Q. In addition, we adopt the convention that the absence of an assertion indicates its negation. Thus all of the above preferences, except for the indifference to contact, have a particularly simple form: a preference for the negation of an assertion over the assertion itself.

When the elementary preferences can be written in this simple form, the induced preference relation on interpretations is given by prioritised subset ordering on the set of assertions made in the various feasible interpretations. As described in §2, we can determine the preference order for any two interpretations by first comparing the assertions made at the highest priority. If the highest priority assertions in one interpretation are a subset of the highest priority assertions in a second interpretation, the first interpretation is preferred. Otherwise,

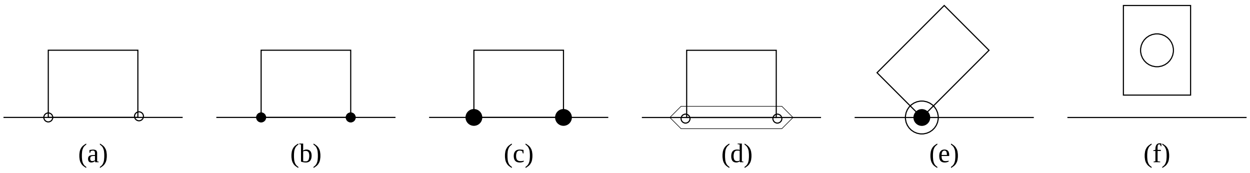

Figure 3: In the presentation of results to follow, we use: (a) small circles to depict sliding contact; (b) small disks for non-sliding contact; (c) large disks for attachment; while (d), (e), and (f) depict LINEARMOTOR(o_1, o_2, c), ANGULARMOTOR(o_1, o_2, p), and BODYMOTOR(o), respectively. For the first two motors, the closed curve surrounds the contact region over which the motors operate, while for body motors the large circle is placed at the object center.

if the two sets of assertions at this priority are not ordered by the subset relation, that is neither set contains the other, then the two interpretations are considered to be unordered. Finally, in the case that the assertions at the highest priority are the same in both interpretations, then we check the assertions at the next lower priority, and so on. This approach, based upon prioritised ordering of elementary preference relations, is similar to prioritised circumscription (Lifschitz, 1985).

To find maximally-preferred models, we search the space of possible interpretations. We perform a breadth-first search, starting with the empty set of assertions, incrementally adding new assertions to this set. Each branch of the search terminates upon finding a minimal set of assertions required for feasible force balancing. Note that because we are indifferent to contacts, we explore every set of admissible contact assertions at each stage of the search. While, in theory, this search could require the testing of every possible interpretation, in practice it often examines only a fraction of the possible interpretations since the search terminates upon finding minimal models.

When the assertions are stratified by a set of priorities, however, we can achieve significant computational savings by performing the search over each priority level separately. For example, under our preference ordering, we can search for minimal sets of motors using only interpretations that contain all admissible attachments. It is critical to note that this algorithm is only correct because of the special structure of the assertions and the domain. The critical property is that if there is a feasible interpretation $i = (C, A)$, and if A' is the set obtained by adding all of the admissible attachments to A, then the interpretation $i = (C, A')$ is also feasible. This property justifies the algorithm above where we set all of the lower priority assertions to the most permissive settings during each stage of the minimization. In general we refer to this property as *monotonicity* (Mann, to appear).

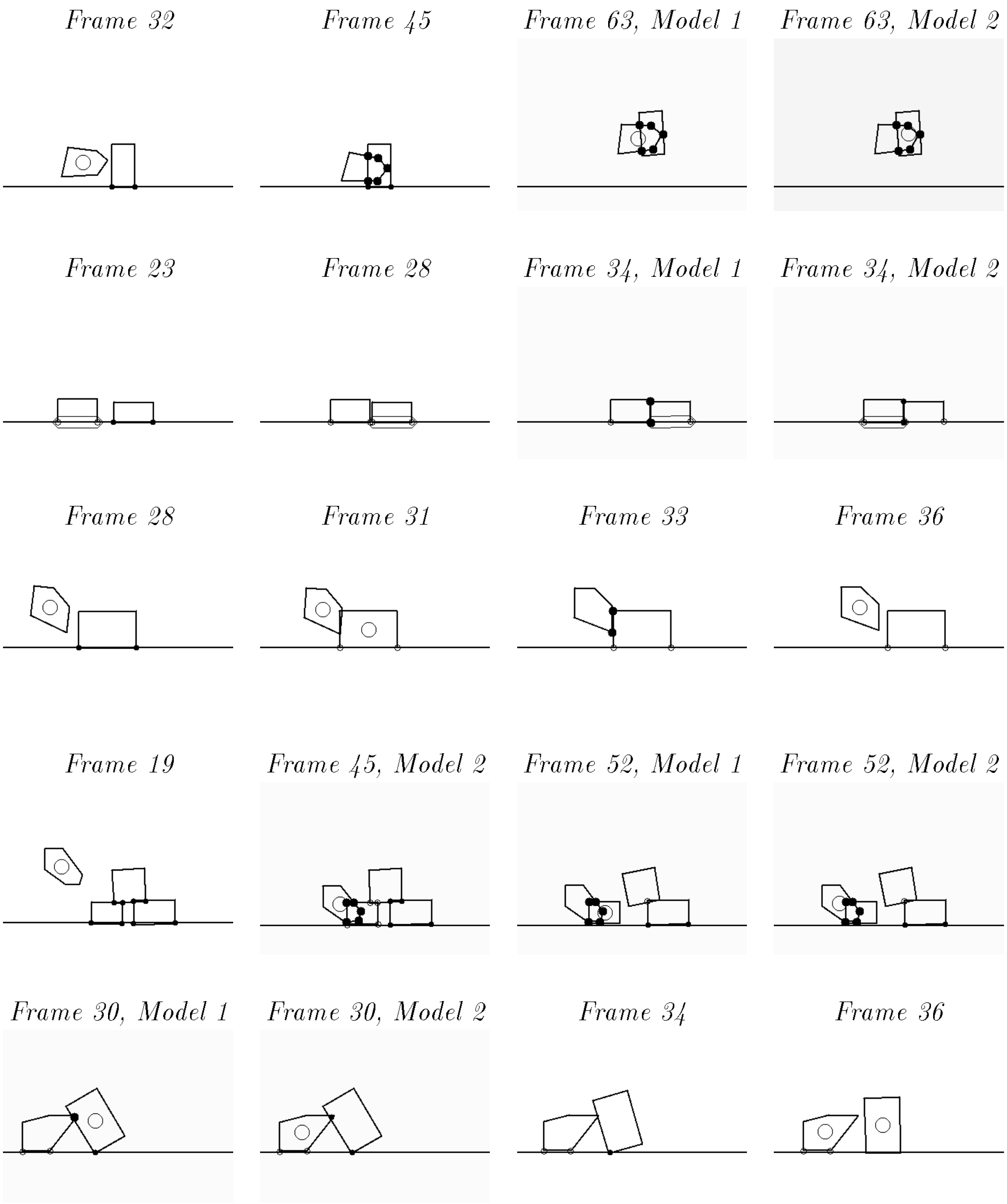

Figure 4: Some preferred models for: coke, cars, hit, arch, and tip. Frames with a non-unique maximally-preferred interpretation are shown with a grey background. Unique interpretations are shown on white.

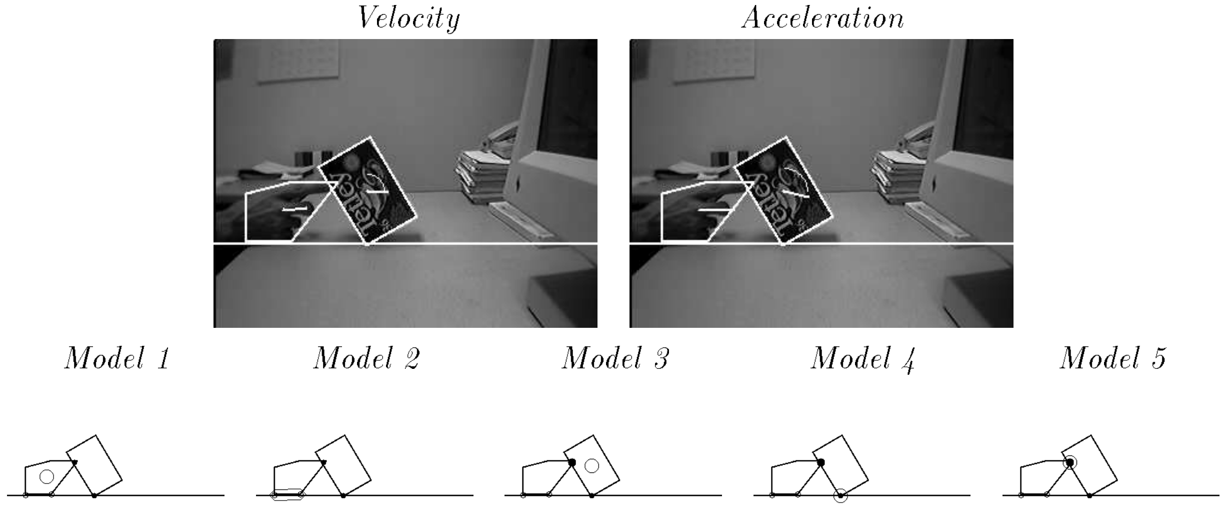

Figure 5: How many ways are there to tip a box? The velocity and acceleration vectors for frame 30 of the tip sequence are depicted on the top row. All five maximally-preferred interpretations are given below. Note that only the first two interpretations are plausible–the last three have a force balance, but are not consistent with rigid body motion. See the text for a discussion.

7 Experimental Results

We have applied our system to several image sequences taken from a desktop environment (see Figure 1). The sequences were taken from a video camera attached to a SunVideo imaging system. MPEG image sequences were acquired at a rate of thirty frames per second and a resolution of 320×240 pixels. The 24-bit colour image sequences were converted to 8-bit grey-scale images used by the tracker.

The input to our system consists of an image sequence, a set of object template images including a polygonal outline of each object, plus an estimate for the position of the objects within the first frame. In addition, we provide an estimate for the *ground plane* which is designated as a ground object in our ontology. Note that the exact shape of the ground plane is not critical, so long as it can be provided to the system in the first frame. Given this input, the tracker provides estimates for the object poses and motions in each frame of the sequence. These estimates, along with the polygonal shapes, are used by the interpretation-construction module.

Figure 4 shows some of the preferred interpretations found for selected frames from each sequence. (Note that the selected frames do not necessarily match those shown in Figure 1.) For each sequence we show frames ordered from left to right. A legend of symbols used to indicate assertions is shown in Figure 3. Note that in order to simplify the presentation, except for the cars sequence, we show only the intepretations involving body motors. For the cars sequence we show only those interpretations using linear motors. While the preferred interpretations are often unique, at times there are multiple interpretations, particularly when objects interact. We highlight frames with multiple preferred interpretations by grey shading.

Our machine interpretations are surprisingly intuitive. For example, the difference between models 1 and 2 in frame 63 of the coke sequence can be interpreted as the hand 'lifting' the can versus the can 'lifting' the hand. Similarly, the difference between models 1 and 2 in frame 34 of the cars sequence can be interpreted as the front car 'pulling' the back car versus the back car 'pushing' the front car. Note that the system correctly hypothesises an attachment between the front and back cars in the 'pulling' interpretation, but does not do so in the 'pushing' interpretation. Without intra-frame analysis or prior information about the objects, all of these interpretations are reasonable. In addition to making inferences about which objects are 'active,' our system can also make inferences about the contact geometry of the scene. For example, during the approach phase of the coke sequence (frame 32), it can infer that the can must be supported by the table.

A physics-based ontology that includes dynamics allows a richer set of descriptions than one based purely on static scenes. This is illustrated by the arch example in the fourth row of Figure 4. The top block of the arch changes from a state of *static* support in frame 19 to *dynamic* tipping in frame 52. In between, in frame 45, the left block is supporting the top block, but the motion of the left block itself must be explained by an active object.[1]

Finally, the tip sequence in the last row of Figure 4 highlights the richness of our dynamic domain. In this sequence, a hand raises a box onto it's corner and allows it to tip to an upright position. Precisely at frame 30, the box changes from a state of being supported by the hand to a state of tipping. There are two interpretations for frame 30 that involve only body motors: one where the hand 'tips' the box and the other where the box 'tips' itself and 'drags' the hand. As with the previous examples, the latter interpretation requires an attachment between the box and the hand while the former does not. Frame 34 shows the situation shortly after the box is tipped. Note that since the center of gravity of the box is ahead of the support point there is significant angular acceleration. Our system correctly infers that this angular acceleration is due to gravity rather than some type of motor.

Figure 5 shows a more detailed analysis of frame 30 where we allow all types of motors. The two images show the object polygons with their velocities and accelerations overlayed on the images. Note that the center of gravity of the box is directly above the pivot point, yet there is significant angular acceleration caused by the hand. Below the images we show all of the preferred interpretations found by our system. The first two interpretations correspond to an active hand 'pushing' the box while the last three show the box 'pulling' the hand. Note that while all interpretations Figure 5 have a force balance, the last three are not consistent with rigid body motion. In particular, it is not *kinematically* feasible for the hand to be both

[1] The second entry in the fourth row of Figure 4 is the second of two possible interpretations for frame 45 of the arch sequence. There should also be a corresponding interpretation in which the left block 'carries' the hand. Our system, however, fails to produce this alternate interpretation. We believe that this is due to a bug in our program and not a fundamental flaw in our approach. We hope to resolve this problem for the final paper.

attached to the box and in edge-to-edge sliding contact with the table. As discussed in §4.2 our system considers only pairwise constraints between objects and does not check for global kinematic consistency. Further tests would be required to rule out these interpretations.

While encouraging, our system exhibits a number of anomalies. These anomalies generally fall into two classes. In the first type of anomaly, slowly-accelerating dynamic objects are sometimes mistaken for static objects, resulting in interpretations where objects with distinct motions are attached in order to reduce the number of active objects. Examples of this are shown in frame 45 of the coke sequence, where the hand attaches to the can, and in frame 33 of the hit sequence, where the hand attaches to the box. Such anomalies are to be expected given that our system only examines single frames in isolation.

A second problem concerns the detection of changing contact relations between objects. In particular, when objects collide, the estimates for relative velocity and acceleration at their contact points differ, resulting in the contact relation being deemed inadmissible. An example of this is shown in frame 28 of the cars sequence where, during a brief interval, the first car is decelerating while the second car is accelerating. While the two cars are actually in contact, this abutment is deemed to be inadmissible due to a large difference in the observed accelerations.[2] Further examples of collisions are shown in frame 31 of the hit sequence and frame 36 of the tip sequence. Again, these anomalies are to be expected since our ontology is restricted to continuous velocity and acceleration, and is not designed to handle impulses and abrupt changes in contact.

8 Related Work

A number of prior systems have represented 'causal' structure using rule-based systems (Fahlman, 1974; Tsotsos, Mylopoulos, Covvey, & Zucker, 1980; Brand, Birnbaum, & Cooper, 1993; Kuniyoshi & Inoue, 1993) and naive physical simulation (Funt, 1980; Joskowicz & Sacks, 1991; Siskind, 1992, 1995). In contrast, our system uses an explicit representation based on Newtonian physics.

A number of other systems have used physically-based representations for scenes in terms of forces in static scenes (Blum et al., 1970), and changing kinematic relations in time-varying scenes (Ikeuchi & Suehiro, 1994; Siskind, 1994). Our system extends these approaches to consider both kinematic and dynamic relations in time-varying scenes containing rigid objects. Shavit and Jepson (1993) present an approach to classifying motion based on the dynamic properties of non-rigid objects.

Our representation of scene properties as logical propositions is related to the approach presented by Reiter and Mackworth (1989). Unlike our system, however, that system considered a simple domain (maps) where feasibility could be expressed as a set of logical constraints. In addition, while that system allowed the representation of uncertainty, preferences

[2]The acceleration of the cars should be equal since they remain in contact after the collision. In our system, however, the interpolator has smoothed over this discontinuity and given unreliable estimates of the acceleration.

were not used.

Finally, while the problem is quite different, our representation of rigid objects borrows heavily from the physical simulation and graphics communities (Featherstone, 1987; Baraff, 1995).

9 Conclusion

In this paper we have presented an implemented computational theory that can derive force-dynamic representations directly from camera input. Our system embodies a rich ontology that includes both kinematic and dynamic properties of the observed objects. Finally, the system provides a representation of uncertainty along with a theory of preferences between multiple interpretations.

While encouraging, this work could be extended in several ways. First, in order to work in a general environment, 3D representations are required. Second, in order to deal with changing contact relations, a theory of *transitions* is required. Such a system would require a treatment of the transfer of momentum, as well as forces, between objects. Finally, in order to describe the causal structure of time-varying scenes, the system will require a representation of object capabilities and how they are expected to change over time.

In summary, we believe that the ability to perform causal analysis lies at the center of human perception and conceptual modeling of the world. Endowing computers with analogous inferential capacity will hopefully allow us to build computers with similar perceptual and computational capabilities in the future.

References

Baraff, D. (1995). Interactive Simulation of Solid Rigid Bodies. *IEEE Computer Graphics and Applications*, *15*(3), 63–75.

Black, M. J., & Jepson, A. D. EigenTracking: Robust Affine Tracking Using a View-Based Representation. Submitted to ECCV–96.

Blum, M., Griffith, A. K., & Neumann, B. (1970). A Stability Test for Configurations of Blocks. A. I. Memo 188, M. I. T. Artificial Intelligence Laboratory.

Brand, M., Birnbaum, L., & Cooper, P. (1993). Sensible Scenes: Visual Understanding of Complex Scenes through Causal Analysis. In *Proceedings of the Eleventh National Conference on Artifical Intelligence*, pp. 588–593.

Fahlman, S. E. (1974). A Planning System for Robot Construction Tasks. *Artificial Intelligence*, *5*(1), 1–49.

Featherstone, R. (1987). *Robot Dynamics Algorithms*. Boston: Kluwer.

Funt, B. V. (1980). Problem-Solving with Diagrammatic Representations. *Artificial Intelligence*, *13*(3), 201–230.

Goldstein, H. (1980). *Classical Mechanics* (second edition). Addison-Wesley.

Ikeuchi, K., & Suehiro, T. (1994). Towards an Assembly Plan from Observation, Part I: Task Recognition with Polyhedral Objects. *IEEE Transactions on Robotics and Automation*, *10*(3), 368–385.

Jenkin, M., & Jepson, A. D. (1994). Detecting Floor Anomalies. In *Proceedings of the British Machine Vision Conference*, pp. 731–740, York, UK.

Jepson, A. D., & Black, M. J. (1993). Mixture Models for Optical Flow. In *Proceedings of the IEEE Conference of Computer Vision and Pattern Recognition*, pp. 760–761.

Jepson, A. D., & Richards, W. (1993). What is a Percept?. Tech. rep. RBCV-TR-93-43, Department of Computer Science, University of Toronto.

Jepson, A. D., Richards, W., & Knill, D. Modal Structure and Reliable Perceptual Inferences. In W. Richards & D. Knill (Eds.), *Perception as Bayesian Inference*. Cambridge University Press.

Joskowicz, L., & Sacks, E. P. (1991). Computational Kinematics. *Artificial Intelligence*, *51*(1–3), 381–416.

Kuniyoshi, Y., & Inoue, H. (1993). Qualitative Recognition of Ongoing Human Action Sequences. In *IJCAI93*, pp. 1600–1609.

Lifschitz, V. (1985). Computing Circumscription. In *IJCAI85*, pp. 121–127.

Luenberger, D. G. (1984). *Linear and Nonlinear Programming*. Addison Wesley.

Mann, R. In preparation.

Reiter, R., & Mackworth, A. (1989). A Logical Framework for Depiction and Image Interpretation. *Artificial Intelligence*, *41*, 125–155.

Richards, W., Jepson, A. D., & Feldman, J. Priors, Preferences and Categorical Percepts. In W. Richards & D. Knill (Eds.), *Perception as Bayesian Inference*. Cambridge University Press.

Shavit, E., & Jepson, A. D. (1993). Qualitative Motion from Visual Dynamics. In *IEEE Workshop on Qualitative Vision*, New York.

Siskind, J. M. (1992). *Naive Physics, Event Perception, Lexical Semantics, and Language Acquisition*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Siskind, J. M. (1994). Axiomatic Support for Event Perception. In P. McKevitt (Ed.), *Proceedings of the AAAI–94 Workshop on the Integration of Natural Language and Vision Processing*, pp. 153–160, Seattle, WA.

Siskind, J. M. (1995). Grounding Language in Perception. *Artificial Intelligence Review*, *8*, 371–391.

Tsotsos, J. K., Mylopoulos, J., Covvey, H. D., & Zucker, S. W. (1980). A Framework for Visual Motion Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *2*(6), 563–573.