Visual Event Classification via Force Dynamics

Jeffrey Mark Siskind
NEC Research Institute, Inc.
4 Independence Way
Princeton NJ 08540 USA
609/951–2705
qobi@research.nj.nec.com
http://www.neci.nj.nec.com/homepages/qobi

Abstract

This paper presents an implemented system, called LEONARD, that classifies simple spatial motion events, such as *pick up* and *put down*, from video input. Unlike previous systems that classify events based on their motion profile, LEONARD uses changes in the state of force-dynamic relations, such as support, contact, and attachment, to distinguish between event types. This paper presents an overview of the entire system, along with the details of the algorithm that recovers force-dynamic interpretations using prioritized circumscription and a stability test based on a reduction to linear programming. This paper also presents an example illustrating the end-to-end performance of LEONARD classifying an event from video input.

Introduction

People can describe what they see. If someone were to pick up a block and ask you what you saw, you could say *The person picked up the block.* In doing so, you describe both *objects*, like people and blocks, and *events*, like pickings up. Most recognition research in machine vision has focussed on recognising objects. In contrast, this paper describes a system for recognising events. Objects correspond roughly to the noun vocabulary in language. In contrast, events correspond roughly to the verb vocabulary in language. The overall goal of this research is to ground the lexical semantics of verbs in visual perception.

A number of reported systems can classify event occurrences from video or simulated video, among them, Yamoto, Ohya, & Ishii (1992), Regier (1992), Pinhanez & Bobick (1995), Starner (1995), Siskind & Morris (1996), Bailey *et al.* (1998), and Bobick & Ivanov (1998). While they differ in their details, by and large, these system classify event occurrences by their motion profile. For example, a *pick up* event is described as a sequence of two subevents: the agent moving towards the patient while the patient is at rest above the source, followed by the agent moving with the patient away from the source. Such systems use some combination of relative and absolute; linear and angular; positions, velocities, and accelerations as the features that drive classification.

These systems follow the tradition of linguists and cognitive scientists, such as Leech (1969), Miller (1972), Schank (1973), Jackendoff (1983), or Pinker (1989), that represent the lexical semantics of verbs via the causal, aspectual, and directional qualities of motion. Some linguists and cognitive scientists, such as Herskovits (1986) and Jackendoff & Landau (1991), have argued that force-dynamic relations (Talmy 1988), such as support, contact, and attachment, are crucial for representing the lexical semantics of spatial prepositions. For example, in some situations, part of what it means for one object to be *on* another object is for the former to be in contact with, and supported by, the latter. In other situations, something can be on something else by way of attachment, as in *the knob on the door*. Siskind (1992) has argued that changes in the state of force-dynamic relations plays a more central role in specifying the lexical semantics of simple spatial motion verbs than motion profile. The particular linear and angular velocities and accelerations don't matter when picking something up or putting something down. What matters is a state change. When picking something up, the patient is initially supported by being on top of the source. Subsequently, the patient is supported by being attached to the agent. Likewise, when putting something down, the reverse is true. The patient starts out being supported by being attached to the agent. It is subsequently supported by being on top of the goal. Furthermore, what distinguishes putting something down from dropping it is that, in the former, the patient is always supported, while in the latter, the patient undergoes unsupported motion.

Siskind (1995), among others, describes a system for recovering force-dynamic relations from simulated video and using those relations to perform event classification. Mann, Jepson, & Siskind (1997), among others, describes a system for recovering force-dynamic relations from video but does not use those relations to perform event classification. This paper describes a system, called LEONARD, that recovers force-dynamic relations from video and uses those relations to perform event classification. It is the first reported system that goes all the way from video to event classification using recovered force dynamics. LEONARD is a complex, comprehensive system. Video input is processed using a real-time colour- and motion-based segmentation procedure to place a convex polygon around each participant object in each input frame. A tracking procedure then computes the corre-

spondence between the polygons in each frame and those in adjacent frames. LEONARD then constructs force-dynamic interpretations of the resulting polygon movie. These interpretations are constructed out of predicates that describe the attachment relations between objects, the qualitative depth of objects, and their groundedness. Some interpretations are consistent in that they describe stable scenes. Others are inconsistent in that they describe unstable scenes. LEONARD performs model reconstruction, selecting as models, only those interpretations that explain the stability of the scene. Kinematic stability analysis is performed efficiently via a reduction to linear programming. There are usually multiple models, i.e. stable interpretations of each scene. LEONARD selects a preferred subset of models using prioritized, cardinality, and temporal circumscription. Event classification is efficiently performed on this preferred subset of models using an interval-based event logic. A precise description of the entire system is beyond the scope of this paper. The remainder of this paper focuses on kinematic stability analysis and model reconstruction. It also presents an example of the entire system in operation. Future papers will describe other components of this system in greater detail.

Kinematic Stability Analysis

Let us consider a simplified world that consists of line segments. Polygons can be treated as collections of rigidly attached line segments. Let us denote line segments by the symbol l. In this simplified world, some line segments will not need to be supported. Such line segments are said to be *grounded*. Let us denote the fact that l is grounded by the property $g(l)$. In this simplified world, the table top and the agent's hand will be grounded.

In this simplified world, line segments can be *joined* together. If l_i and l_j are joined, the constraint on their relative motion is specified by three relations $\leftrightarrow_1$, $\leftrightarrow_2$, and $\leftrightarrow_\theta$. If $l_i \leftrightarrow_1 l_j$, then the position of the joint along l_i is fixed. Likewise, if $l_i \leftrightarrow_2 l_j$, then the position of the joint along l_j is fixed. And if $l_i \leftrightarrow_\theta l_j$, then the relative orientation of l_i and l_j is fixed. Combinations of these three relations allow specifying a variety of joint types. If $l_i \leftrightarrow_1 l_j \wedge l_i \leftrightarrow_2 l_j \wedge l_i \leftrightarrow_\theta l_j$, then l_i and l_j are rigidly joined. If $l_i \leftrightarrow_1 l_j \wedge l_i \leftrightarrow_2 l_j \wedge l_i \not\leftrightarrow_\theta l_j$, then l_i and l_j are joined by a revolute joint. If $l_i \not\leftrightarrow_1 l_j \wedge l_i \leftrightarrow_2 l_j \wedge l_i \leftrightarrow_\theta l_j$, then l_i and l_j are joined by a prismatic joint that allows l_j to slide along l_i. If $l_i \leftrightarrow_1 l_j \wedge l_i \not\leftrightarrow_2 l_j \wedge l_i \leftrightarrow_\theta l_j$, then l_i and l_j are joined by a prismatic joint that allows l_i to slide along l_j. If $l_i \not\leftrightarrow_1 l_j \wedge l_i \not\leftrightarrow_2 l_j \wedge l_i \not\leftrightarrow_\theta l_j$, then l_i and l_j are not joined. A total of eight different kinds of joints are possible, including ones that are simultaneously revolute and prismatic.

In this simplified world, line segments reside on parallel planes that are perpendicular to the focal axis of the observer. This simplified world uses an impoverished notion of depth. All that is important is whether two given line segments reside on the same plane. Such line segments are said to be on the *same layer*. Let us denote the fact that l_i and l_j are on the same layer by the relation $l_i \bowtie l_j$. This impoverished notion of depth lacks any notion of depth order. It cannot model objects being in front of or behind other objects. It also lacks any notion of adjacency in depth. It cannot model objects touching one another along the focal axis of the observer.

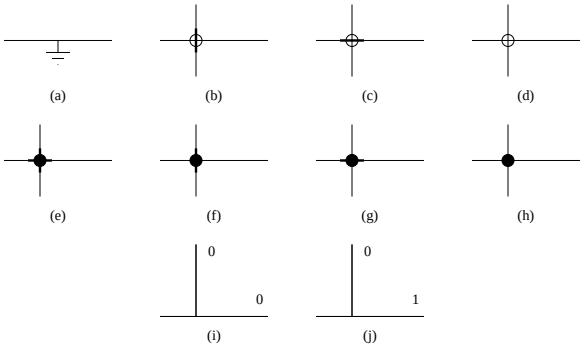

Figure 1: A graphical representation of scene interpretations. Consider the vertical lines to be l_i and the horizontal lines to be l_j. (a) depicts $g(l_j)$. (b) depicts $l_i \not\leftrightarrow_1 l_j \wedge l_i \leftrightarrow_2 l_j \wedge l_i \not\leftrightarrow_\theta l_j$. (c) depicts $l_i \leftrightarrow_1 l_j \wedge l_i \not\leftrightarrow_2 l_j \wedge l_i \not\leftrightarrow_\theta l_j$. (d) depicts $l_i \leftrightarrow_1 l_j \wedge l_i \leftrightarrow_2 l_j \wedge l_i \not\leftrightarrow_\theta l_j$. (e) depicts $l_i \not\leftrightarrow_1 l_j \wedge l_i \not\leftrightarrow_2 l_j \wedge l_i \leftrightarrow_\theta l_j$. (f) depicts $l_i \not\leftrightarrow_1 l_j \wedge l_i \leftrightarrow_2 l_j \wedge l_i \leftrightarrow_\theta l_j$. (g) depicts $l_i \leftrightarrow_1 l_j \wedge l_i \not\leftrightarrow_2 l_j \wedge l_i \leftrightarrow_\theta l_j$. (h) depicts $l_i \leftrightarrow_1 l_j \wedge l_i \leftrightarrow_2 l_j \wedge l_i \leftrightarrow_\theta l_j$. The $\bowtie$ relation is depicted by assigning layer indices to line segments. (i) depicts $l_i \bowtie l_j$. (j) depicts $l_i \not\bowtie l_j$.

A *scene* is a set L of line segments. An *interpretation* of scene is a quintuple $\langle g, \leftrightarrow_1, \leftrightarrow_2, \leftrightarrow_\theta, \bowtie \rangle$. It is convenient to depict scene interpretations graphically. Figure 1 shows the graphical representation of the predicates g, $\leftrightarrow_1$, $\leftrightarrow_2$, $\leftrightarrow_\theta$, and $\bowtie$.

An interpretation is *admissible* if the following conditions hold:

- For all l_i and l_j, if $l_i \leftrightarrow_1 l_j$, $l_i \leftrightarrow_2 l_j$, or $l_i \leftrightarrow_\theta l_j$, then l_i intersects l_j. In other words, attached line segments must intersect.
- For all l_i and l_j, $l_i \leftrightarrow_1 l_j$ iff $l_j \leftrightarrow_2 l_i$.
- $\leftrightarrow_\theta$ is symmetric.
- For all l_i and l_j, if $l_i \bowtie l_j$, then l_i and l_j do not overlap. In other words, line segments on the same layer must not overlap. Two line segments overlap if they intersect in a noncollinear fashion and the point of intersection is not an endpoint of either line segment.
- $\bowtie$ is symmetric and transitive.

(Note that whether two line segments intersect or overlap is purely a geometric property of the scene L and is independent of any interpretation of that scene.) Only admissible interpretations will be considered.

Statically, the line segments in a scene have fixed positions and orientations. Let us denote the coordinates of a point p as $x(p)$ and $y(p)$. And let us denote the endpoints of a line segment l as $p(l)$ and $q(l)$. And let us denote the length of a line segment l as $||l||$. And let us denote the position of a line segment l as the position of its midpoint $c(l)$. And let us denote the orientation $\theta(l)$ of a line segment l as the angle of the vector from $p(l)$ to $q(l)$. The quantities $p(l)$, $q(l)$,

$||l||$, $c(l)$, and $\theta(l)$ are all fixed given a static scene. And $p(l)$ and $q(l)$ are related to $||l||$, $c(l)$, and $\theta(l)$ as follows:

$$\begin{aligned} c(l) &= \frac{p(l)+q(l)}{2} \\ ||l|| &= \sqrt{(q(l)-p(l))\cdot(q(l)-p(l))} \\ \theta(l) &= \tan^{-1}\frac{y(q(l))-y(p(l))}{x(q(l))-x(p(l))} \\ x(p(l)) &= x(c(l))-\frac{1}{2}||l||\cos\theta(l) \\ y(p(l)) &= y(c(l))-\frac{1}{2}||l||\sin\theta(l) \\ x(q(l)) &= x(c(l))+\frac{1}{2}||l||\cos\theta(l) \\ y(q(l)) &= y(c(l))+\frac{1}{2}||l||\sin\theta(l) \end{aligned}$$

Let us postulate an unknown instantaneous motion for each line segment in the scene. This can be represented by associating a linear and angular velocity with each line segment. Let us denote such velocities with the variables $\dot{c}(l)$ and $\dot{\theta}(l)$. Let us assume that there is no motion in depth so the $\bowtie$ relation does not change. If the scene contains n line segments, then there will be $3n$ scalar variables, because $\dot{c}$ has x and y components. Assuming that the line segments are rigid, i.e. that instantaneous motion does not lead to a change in their length, one can relate $\dot{p}(l)$ and $\dot{q}(l)$, the instantaneous velocities of the endpoints, to $\dot{c}(l)$ and $\dot{\theta}(l)$, using the chain rule as follows:

$$\begin{aligned} \dot{p}(l) &= \frac{\partial p(l)}{\partial x(c(l))}x(\dot{c}(l))+\frac{\partial p(l)}{\partial y(c(l))}y(\dot{c}(l))+\frac{\partial p(l)}{\partial \theta(l)}\dot{\theta}(l) \\ \dot{q}(l) &= \frac{\partial q(l)}{\partial x(c(l))}x(\dot{c}(l))+\frac{\partial q(l)}{\partial y(c(l))}y(\dot{c}(l))+\frac{\partial q(l)}{\partial \theta(l)}\dot{\theta}(l) \end{aligned}$$

Note that $\dot{p}(l)$ and $\dot{q}(l)$ are linear in $\dot{c}(l)$ and $\dot{\theta}(l)$.

Each of the components of an admissible interpretation of a scene can be viewed as imposing constraints on the instantaneous motions of the line segments in that scene. The simplest case is the g property. If $g(l)$, then $\dot{c}(l)=0$ and $\dot{\theta}(l)=0$. Note that these equations are linear in $\dot{c}(l)$ and $\dot{\theta}(l)$.

Let us now consider the $l_i \leftrightarrow_1 l_j$ and $l_i \leftrightarrow_2 l_j$ relations and the constraint that they imposes on the motions of l_i and l_j. First, let us denote the intersection of l_i and l_j as $I(l_i,l_j)$. If we let

$$\begin{aligned} A &= \begin{pmatrix} y(p(l_i))-y(q(l_i)) & x(q(l_i))-x(p(l_i)) \\ y(p(l_j))-y(q(l_j)) & x(q(l_j))-x(p(l_j)) \end{pmatrix} \\ b &= \begin{pmatrix} y(p(l_i))(x(q(l_i))-x(p(l_i))) \\ +x(p(l_i))(y(p(l_i))-y(q(l_i))) \\ y(p(l_j))(x(q(l_j))-x(p(l_j))) \\ +x(p(l_j))(y(p(l_j))-y(q(l_j))) \end{pmatrix} \end{aligned}$$

then $I(l_i,l_j)=A^{-1}b$.

Next, let us compute $\dot{I}(l_i,l_j)$, the velocity of the intersection of the two lines segments l_i and l_j as they move. Let α be the vector containing the elements $x(p(l_i))$, $y(p(l_i))$, $x(q(l_i))$, $y(q(l_i))$, $x(p(l_j))$, $y(p(l_j))$, $x(q(l_j))$, and $y(q(l_j))$. And let β be the vector containing the elements $x(c(l_i))$, $y(c(l_i))$, $\theta(l_i)$, $x(c(l_j))$, $y(c(l_j))$, and $\theta(l_j)$. And let γ be the vector where $\gamma_k = \frac{\partial I}{\partial \alpha_k}$. And let D be the matrix where $D_{kl} = \frac{\partial \alpha_k}{\partial \beta_l}$. $\dot{I}(l_i,l_j)$ can be computed by the chain rule as follows:

$$\dot{I}(l_i,l_j) = \gamma^T D\dot{\beta}$$

Note that $\dot{I}(l_i,l_j)$ is linear in $\dot{c}(l_i)$, $\dot{\theta}(l_i)$, $\dot{c}(l_j)$, and $\dot{\theta}(l_j)$ because all of the partial derivatives are constant.

Next, let us denote by $\rho(p,l)$, where p is a point on l, the fraction of the distance where p lies between $p(l)$ and $q(l)$.

$$\rho(p,l) = \begin{cases} \frac{y(p)-y(p(l))}{y(q(l))-y(p(l))} & x(p(l))=x(q(l)) \\ \frac{x(p)-x(p(l))}{x(q(l))-x(p(l))} & \text{otherwise} \end{cases}$$

And if $0 \le \rho \le 1$, let us denote by $l(\rho)$ the point that is the fraction ρ of the distance between $p(l)$ and $q(l)$.

$$l(\rho) = p(l) + \rho(q(l)-p(l))$$

And let us denote by $\dot{l}(\rho)$ the velocity of the point that is the fraction ρ of the distance between $p(l)$ and $q(l)$ as l moves. Let α be the vector containing the elements $x(p(l))$, $y(p(l))$, $x(q(l))$, and $y(q(l))$. And let β be the vector containing the elements $x(c(l))$, $y(c(l))$, and $\theta(l)$. And let γ be the vector where $\gamma_k = \frac{\partial l(\rho)}{\partial \alpha_k}$. And let D be the matrix where $D_{kl} = \frac{\partial \alpha_k}{\partial \beta_l}$. Again, by the chain rule:

$$\dot{l}(\rho) = \gamma^T D\dot{\beta}$$

Again, note that $\dot{l}(\rho)$ is linear in $\dot{c}(l)$ and $\dot{\theta}(l)$ because all of the partial derivatives are constant.

The $\leftrightarrow_1$ constraint can now be formulated as follows: if $l_i \leftrightarrow_1 l_j$, then

$$\dot{l}_i(\rho(I(l_i,l_j),l_i)) = \dot{I}(l_i,l_j)$$

And the $\leftrightarrow_2$ constraint can now be formulated as follows: if $l_i \leftrightarrow_2 l_j$, then

$$\dot{l}_j(\rho(I(l_i,l_j),l_j)) = \dot{I}(l_i,l_j)$$

Again, note that these equations are linear in $\dot{c}(l_i)$, $\dot{\theta}(l_i)$, $\dot{c}(l_j)$, and $\dot{\theta}(l_j)$.

Let us now consider the $l_i \leftrightarrow_\theta l_j$ relation and the constraint that it imposes on the motions of l_i and l_j. If $l_i \leftrightarrow_\theta l_j$, then

$$\dot{\theta}(l_i) = \dot{\theta}(l_j)$$

Again, note that this equation is linear in $\dot{\theta}(l_i)$ and $\dot{\theta}(l_j)$.

The same-layer relation $l_i \bowtie l_j$ imposes the constraint that the motion of l_i and l_j must not lead to an instantaneous penetration of one by the other. An instantaneous penetration can occur only when the endpoint of one line segment touches the other line segment. Without loss of generality, let us assume that $p(l_i)$ touches l_j. Let $\overline{p}$ denote a vector of the same magnitude as p rotated counterclockwise $90°$.

$$\overline{(x,y)} = (-y,x)$$

Let σ be a vector that is normal to l_j, in the direction towards l_i.

$$\sigma = -[\overline{q(l_j) - p(l_j)} \cdot (q(l_i) - p(l_i))]\overline{q(l_j) - p(l_j)}$$

An instantaneous penetration can occur only when the velocity of $p(l_i)$ in the direction of σ is less than the velocity of the point of contact in the same direction. The velocity of $p(l_i)$ is $\dot{p}(l_i)$. And the velocity of the point of contact is $\dot{l}_j(\rho(p(l_i), l_j))$. Thus if $l_i \bowtie l_j$ and $p(l_i)$ touches l_j, then

$$\dot{p}(l_i) \cdot \sigma \leq \dot{l}_j(\rho(p(l_i), l_j)) \cdot \sigma \quad (1)$$

Again, note that this inequality is linear in $\dot{c}(l_i)$, $\dot{\theta}(l_i)$, $\dot{c}(l_j)$, and $\dot{\theta}(l_j)$.

We wish to determine the stability of a scene under an admissible interpretation. A scene is unstable if there is an assignment of linear and angular velocities to the line segments in the scene that satisfies the above constraints and decreases the potential energy of the scene. The potential energy of a scene is the sum of the potential energies of the line segments in that scene. The potential energy of a line segment l is proportional to its mass times $y(c(l))$. We can take the mass of a line segment to be proportional to its length. So the potential energy E can be taken as $\sum_{l \in L} ||l|| y(c(l))$. The potential energy can decrease if $\dot{E} < 0$. By scale invariance, if $\dot{E}$ can be less than zero, then it can be equal to any value less than zero, in particular -1. Thus a scene is unstable under an admissible interpretation iff the constraint $\dot{E} = -1$ is consistent with the above constraints. Note that $\dot{E}$ is linear in all of the $\dot{c}(l)$ values. Thus the stability of a scene under an admissible interpretation can be determined by a reduction to linear programming.

Model Reconstruction

Let us define a *model* of a scene as an admissible interpretation under which the scene is stable. LEONARD enumerates the models of each frame in each movie it processes. This is called *model reconstruction*. There are usually multiple models of any given scene. For example, if the scene contains a collection of overlapping polygons, then the polygons can be rigidly attached and any of them can be grounded. In a certain sense, some models make weaker assumptions than others. For example, it is always possible to explain the stability of an object by grounding it. Thus a model that has fewer grounded objects makes weaker assumptions than a model with more grounded objects. Similarly, whenever it is possible to explain the stability of an object that is supported by being above and on the same layer as another stable object, it is also possible to explain its stability by instead being attached to that object. Thus a model that has fewer attachment relations makes weaker assumptions than a model with more attachment relations. Accordingly, during model reconstruction, LEONARD selects models that make the weakest assumptions. It does so by a process of prioritized circumscription (McCarthy 1980).

To limit the search space, LEONARD considers only rigid and revolute joints. It does not consider prismatic joints. In other words, only interpretations that meet the following constraint are considered:

$$(\forall l_i, l_j \in L) \left\{ \begin{array}{l} [(l_i \leftrightarrow_1 l_i) \leftrightarrow (l_i \leftrightarrow_2 l_i)] \wedge \\ [(l_i \leftrightarrow_\theta l_i) \rightarrow (l_i \leftrightarrow_1 l_i)] \end{array} \right\}$$

Let us define several preference relations between components of interpretations. First, let us define a preference relation between two grounded properties. Let us say that g is preferred to g', denoted $g \prec g'$, if $g \neq g'$ and $(\forall l \in L) g(l) \rightarrow g'(l)$. In other words, g is preferred to g' if they are different and every grounded line segment in the former is grounded in the latter. Along these lines, let us define a similar preference relation between two $\leftrightarrow_1$ relations. Let us say that $\leftrightarrow_1$ is preferred to $\leftrightarrow_1'$, denoted $\leftrightarrow_1 \prec \leftrightarrow_1'$, if $\leftrightarrow_1 \neq \leftrightarrow_1'$ and $(\forall l_i, l_j \in L) l_i \leftrightarrow_1 l_j \rightarrow l_i \leftrightarrow_1' l_j$. In other words, $\leftrightarrow_1$ is preferred to $\leftrightarrow_1'$ if they are different and every pair of line segments that is attached in the former is attached in the latter. Similarly, let us define a preference relation between two $\leftrightarrow_\theta$ relations. Let us say that $\leftrightarrow_\theta$ is preferred to $\leftrightarrow_\theta'$, denoted $\leftrightarrow_\theta \prec \leftrightarrow_\theta'$, if $\leftrightarrow_\theta \neq \leftrightarrow_\theta'$ and $(\forall l_i, l_j \in L) l_i \leftrightarrow_\theta l_j \rightarrow l_i \leftrightarrow_\theta' l_j$. In other words, $\leftrightarrow_\theta$ is preferred to $\leftrightarrow_\theta'$ if they are different and every pair of line segments that is rigidly attached in the former is rigidly attached in the latter. Finally, let us define a preference relation between two same-layer relations. Let us say that $\bowtie$ is preferred to $\bowtie'$, denoted $\bowtie \prec \bowtie'$, if $\bowtie \neq \bowtie'$ and $(\forall l_i, l_j \in L) l_i \bowtie l_j \rightarrow l_i \bowtie' l_j$. In other words, $\bowtie$ is preferred to $\bowtie'$ if they differ and every pair of line segments that is on the same layer in the former is on the same layer in the latter.

Given a set G of grounded properties, its minimal elements $\hat{g}$ are those grounded properties $g \in G$ such that there is no $g' \in G$ where $g' \prec g$. Similarly for $\leftrightarrow_1$, $\leftrightarrow_\theta$, and $\bowtie$.

LEONARD uses the following prioritized circumscription process. It first finds all minimal grounded properties $\hat{g}$ such that there exist relations $\leftrightarrow_1$, $\leftrightarrow_\theta$, and $\bowtie$ where the interpretation $\langle \hat{g}, \leftrightarrow_1, \leftrightarrow_2, \leftrightarrow_\theta, \bowtie \rangle$ is admissible and the scene is stable under that interpretation. Note that there must be at least one such minimal grounded property $\hat{g}$. For each such minimal grounded property $\hat{g}$, it then finds all minimal $\widehat{\leftrightarrow_1}$ relations such that there exist relations $\leftrightarrow_\theta$ and $\bowtie$ where the interpretation $\langle \hat{g}, \widehat{\leftrightarrow_1}, \widehat{\leftrightarrow_1}, \leftrightarrow_\theta, \bowtie \rangle$ is admissible and the scene is stable under that interpretation. Note that there must be at least one such minimal $\widehat{\leftrightarrow_1}$ relation for each minimal grounded property. For each such minimal $\widehat{\leftrightarrow_1}$ relation, taken with the corresponding minimal grounded property $\hat{g}$, it then finds all minimal $\widehat{\leftrightarrow_\theta}$ relations such that there exists a same-layer relation $\bowtie$ where the interpretation $\langle \hat{g}, \widehat{\leftrightarrow_1}, \widehat{\leftrightarrow_1}, \widehat{\leftrightarrow_\theta}, \bowtie \rangle$ is admissible and the scene is stable under that interpretation. Note that there must be at least one such minimal $\widehat{\leftrightarrow_\theta}$ relation for each minimal minimal $\widehat{\leftrightarrow_1}$ relation. Finally, for each such minimal $\widehat{\leftrightarrow_\theta}$ relation, taken with the corresponding minimal $\widehat{\leftrightarrow_1}$ relation and minimal grounded property $\hat{g}$, it then finds all minimal same-layer relations $\widehat{\bowtie}$ such that the interpretation $\langle \hat{g}, \widehat{\leftrightarrow_1}, \widehat{\leftrightarrow_1}, \widehat{\leftrightarrow_\theta}, \widehat{\bowtie} \rangle$ is admissible and the scene is stable under that interpretation. Note that there must be at least one such minimal same-layer relation $\widehat{\bowtie}$ for each such minimal $\widehat{\leftrightarrow_\theta}$ relation. For each such minimal same-layer relation $\widehat{\bowtie}$, taken with the correspond-

ing minimal $\widehat{\leftrightarrow_\theta}$ and $\widehat{\leftrightarrow_1}$ relations and minimal grounded property $\hat{g}$, prioritized circumscription returns the interpretation $\langle \hat{g}, \widehat{\leftrightarrow_1}, \widehat{\leftrightarrow_1}, \widehat{\leftrightarrow_\theta}, \widehat{\bowtie} \rangle$ as a minimal model of the scene.

The above prioritized circumscription procedure orders models by an inclusion relation and selects minimal models according to that inclusion relation. This has the following disadvantage. If a scene has one block resting on top of another block, there will be two minimal models: one where the bottom block is grounded and the top block is on the same layer as the bottom block and one where the top block is grounded and the bottom block is attached to the top block. The latter has more attachment assertions than the former but is still minimal because it is generated for a different minimal grounded property $\hat{g}$. Nonetheless, it is desirable to prefer the former model to the latter model. This is done by a second pass circumscription that uses a cardinality-based preference metric rather than one based on inclusion. Let us define the cardinality of a grounded property g, denoted $||g||$, as the number of line segments $l \in L$ for which $g(l)$ is true. And let us define the cardinality of a $\leftrightarrow_1$ relation, denoted $|| \leftrightarrow_1 ||$ as the number of pairs of line segments $l_i, l_j \in L$ for which $l_i \leftrightarrow_1 l_j$ is true. Similarly for the $\leftrightarrow_\theta$ and $\bowtie$ relations. Furthermore, let us define the cardinality of an interpretation I, denoted $||I||$, as the quadruple $\langle ||g||, || \leftrightarrow_1 ||, || \leftrightarrow_\theta ||, || \bowtie || \rangle$. Now, let us define a preference relation between two interpretations. Let us say that I is preferred to I', denoted $I \prec I'$, if $||I||$ is lexicographically less than $||I'||$. Given a set $\mathcal{I}$ of interpretation, its minimal elements $\hat{I}$ are those interpretations $I \in \mathcal{I}$ such that there is no $I' \in \mathcal{I}$ where $I' \prec I$. Cardinality circumscription returns the minimal elements of the set of minimal models produced by prioritized circumscription.

Prioritized and cardinality circumscription are not sufficient to prune all of the spurious models. The following situation often arises. During a *pick up* event, the agent is grounded before it grasps the patient but once the patient is grasped and lifted there is an ambiguity as to whether the agent is grounded and the patient is supported by being attached to the agent or whether the patient is grounded and the agent is supported by being attached to the patient. In some sense, the former is preferable to the later since it does not require the ground assertion to move from the agent to the patient. More generally, sequences of models are preferred when they entail fewer changes in assertions. This is a form of temporal circumscription (Mann & Jepson 1998). Taken together, prioritized, cardinality, and temporal circumscription typically yield a small number of models for each movie that usually correspond to natural pretheoretic human intuition.

Examples

The techniques described in this paper have been implemented in a system called LEONARD. Figure 2 shows the results of processing four short movies with LEONARD. Each column shows a subset of the frames from a single movie. From left to right, the movies have 29, 34, 16, and 16 frames respectively. Each movie was processed by the segmentation procedure to place a convex polygon around the coloured and moving objects in each frame. The tracking procedure computed the correspondence between the polygons in each frame and those in the adjacent frames. The model reconstruction procedure was used to construct force-dynamic models of each frame. Prioritized, cardinality, and temporal circumscription were used to prune the space of models. Each frame is shown with the results of segmentation and model reconstruction superimposed on the original video image.

For the leftmost column, notice that LEONARD determines that the lower block and the hand are grounded for the entire movie, that the upper block is supported by being on the same layer as the lower block for frames 2, 4, and 8, and that the upper block is supported by being rigidly attached to the hand for frames 20, 22, and 24. For the second column, notice that LEONARD determines that the lower block and the hand are grounded for the entire movie, that the upper block is supported by being rigidly attached to the hand for frames 5 and 10, and that the upper block is supported by being on the same layer as the lower block for frames 21, 24, 26, and 29. For the third column, notice that LEONARD determines that the lower block and the hand are grounded for the entire movie and that the upper block is supported for the entire movie by being on the same layer as the lower block. For the rightmost column, notice that LEONARD determines that the lower block and the hand are grounded for the entire movie and that the upper block is supported for the entire movie by being rigidly attached to the hand.

LEONARD was given the following lexicon when processing these four movies:

$$\textsc{PickUp}(x,y) \triangleq \left[\begin{array}{l} \neg\textsc{Supported}(x) \wedge \\ \textsc{Supported}(y) \wedge \\ \left(\begin{array}{l} \neg\textsc{Attached}(x,y); \\ \textsc{Attached}(x,y) \end{array}\right) \end{array}\right]$$

$$\textsc{PutDown}(x,y) \triangleq \left[\begin{array}{l} \neg\textsc{Supported}(x) \wedge \\ \textsc{Supported}(y) \wedge \\ \left(\begin{array}{l} \textsc{Attached}(x,y); \\ \neg\textsc{Attached}(x,y) \end{array}\right) \end{array}\right]$$

Essentially, these define *pick up* and *put down* as events where the agent is not supported throughout the event, the patient is supported throughout the event, and the agent grasps or releases the patient respectively. LEONARD correctly recognises the movie in the leftmost column as depicting a *pick up* event and the movie in the second column as depicting a *put down* event. More importantly, LEONARD correctly recognises that the remaining two movies do not depict any of the defined event types. Note that systems that classify events based on motion profiles will often mistakingly classify these last two movies as either *pick up* or *put down* events because they have similar motion profiles.

Conclusion

I have presented a comprehensive system that recovers force-dynamic interpretations from video and uses those interpretations to recognise event occurrences. Force dynamics is fundamentally more robust than motion profile for classifying events. Two occurrences of the same event type

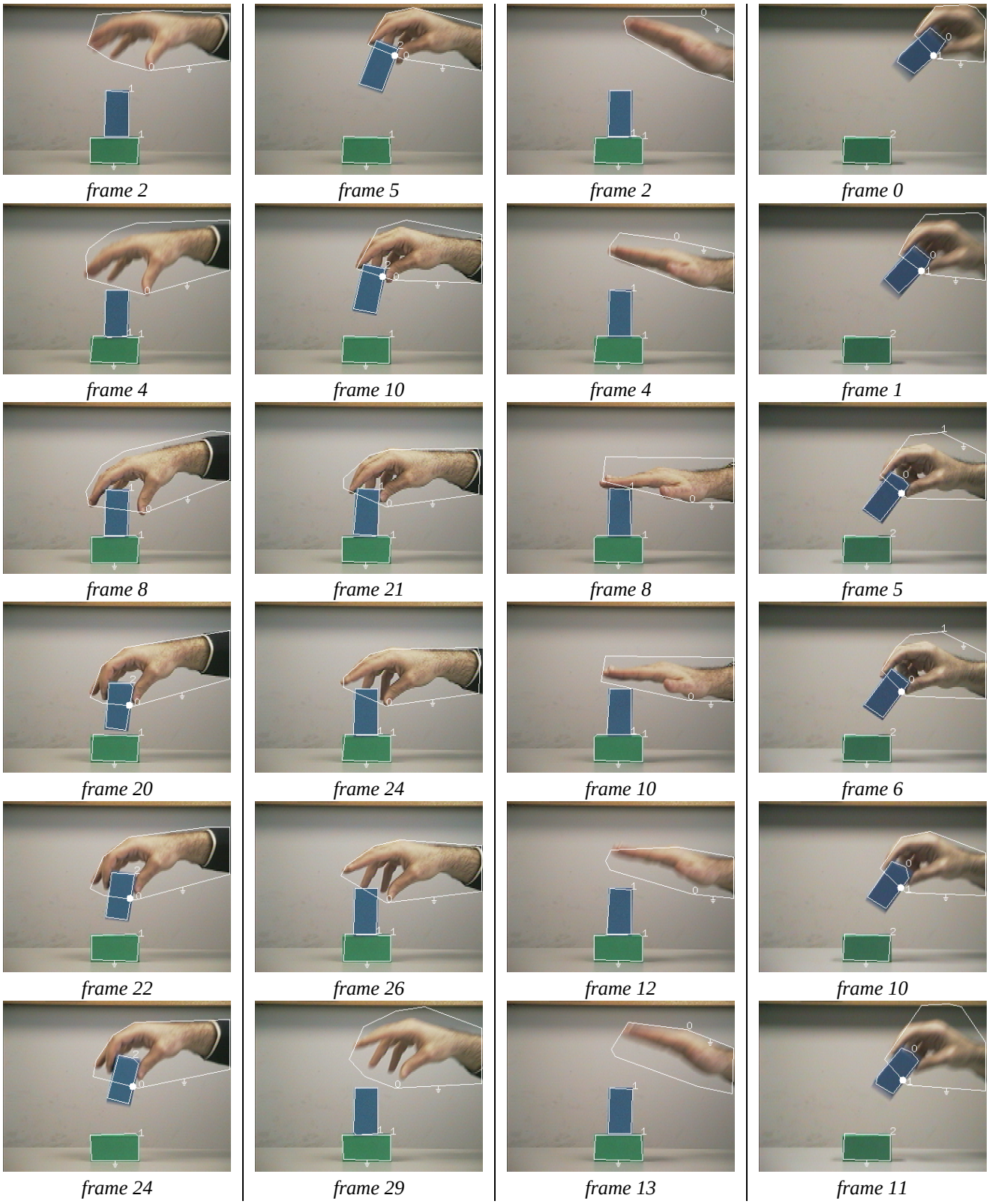

Figure 2: The results of processing four short movies with LEONARD. The segmented polygons and force-dynamic models for each frame are overlayed on the video image for that frame. LEONARD successfully recognises the movie in the leftmost column as a *pick up* event and the movie in the second column as a *put down* event.

might have very different motion profiles. And occurrences of two different event types might have very similar motion profiles. This has been demonstrated by an implementation that distinguishes between occurrences and nonoccurrences of *pick up* and *put down* events despite similar motion profiles.

Acknowledgments

Amit Roy Chowdhury implemented an early version of the stability-checking algorithm described in this paper.

References

Bailey, D. R.; Chang, N.; Feldman, J.; and Narayanan, S. 1998. Extending embodied lexical development. In *Proceedings of the* 20th *Annual Conference of the Cognitive Science Society*.

Bobick, A. F., and Ivanov, Y. A. 1998. Action recognition using probabilistic parsing. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 196–202.

Herskovits, A. 1986. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English*. New York, NY: Cambridge University Press.

Jackendoff, R., and Landau, B. 1991. Spatial language and spatial cognition. In Napoli, D. J., and Kegl, J. A., eds., *Bridges Between Psychology and Linguistics: A Swarthmore Festschrift for Lila Gleitman*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jackendoff, R. 1983. *Semantics and Cognition*. Cambridge, MA: The MIT Press.

Leech, G. N. 1969. *Towards a Semantic Description of English*. Indiana University Press.

Mann, R., and Jepson, A. 1998. Toward the computational perception of action. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 794–799.

Mann, R.; Jepson, A.; and Siskind, J. M. 1997. The computational perception of scene dynamics. *Computer Vision and Image Understanding* 65(2).

McCarthy, J. 1980. Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence* 13(1–2):27–39.

Miller, G. A. 1972. English verbs of motion: A case study in semantics and lexical memory. In Melton, A. W., and Martin, E., eds., *Coding Processes in Human Memory*. Washington, DC: V. H. Winston and Sons, Inc. chapter 14, 335–372.

Pinhanez, C., and Bobick, A. 1995. Scripts in machine understanding of image sequences. In *AAAI Fall Symposium Series on Computational Models for Integrating Language and Vision*.

Pinker, S. 1989. *Learnability and Cognition*. Cambridge, MA: The MIT Press.

Regier, T. P. 1992. *The Acquisition of Lexical Semantics for Spatial Terms: A Connectionist Model of Perceptual Categorization*. Ph.D. Dissertation, University of California at Berkeley.

Schank, R. C. 1973. The fourteen primitive actions and their inferences. Memo AIM-183, Stanford Artificial Intelligence Laboratory.

Siskind, J. M., and Morris, Q. 1996. A maximum-likelihood approach to visual event classification. In *Proceedings of the Fourth European Conference on Computer Vision*, 347–360. Cambridge, UK: Springer-Verlag.

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

Siskind, J. M. 1995. Grounding language in perception. *Artificial Intelligence Review* 8:371–391.

Starner, T. E. 1995. Visual recognition of american sign language using hidden markov models. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA.

Talmy, L. 1988. Force dynamics in language and cognition. *Cognitive Science* 12:49–100.

Yamoto, J.; Ohya, J.; and Ishii, K. 1992. Recognizing human action in time-sequential images using hidden markov model. In *Proceedings of the 1992 IEEE Conference on Computer Vision and Pattern Recognition*, 379–385. IEEE Press.