

Understanding Manipulation in Video

Matthew Brand
MIT Media Lab

20 Ames St., Cambridge, MA 02139

brand@media.mit.edu

<http://www.media.mit.edu/~brand/>

Abstract

Manipulations are a significant subset of human gestures that are distinguished by the fact that their logic and meaning are particularly clear, being heavily constrained by physical causality. We present techniques and causal semantics for interpreting video of manipulation tasks such as disassembly. Psychologically-based causal constraints are used to detect meaningful changes in the integrity and motions of foreground-segmented blobs; a small causal model of manipulation is used to disambiguate and parse these into a coherent account of video's action. The causal constraints are drawn from studies of infant perceptual development; as with infants, they precede and may possibly even bootstrap the ability to reliably segment still objects. Our implementation produces a script of the causal evolution of the scene—output that supports cartoon summary, automated editing, and higher-level reasoning.

1 Understanding manipulation

Much visual experience is devoted to watching humans manipulate their environment. Manipulations, like gestures, are meaningful sequences of body articulations. However, two factors make manipulations particularly amenable to machine understanding. First, they derive their meaning from a restricted semantics that is ultimately based in the physics of human-scaled objects. Second, they are composed of actions (e.g., picking up) that are defined by particular causal events (e.g., engagement) that in turn have typical visual signatures (e.g., induced motion). This suggests that it may be possible to reconstruct the action plan of a manipulation video by reasoning about the motions and collisions of surfaces. Indeed, it appears that this strategy plays an early and important developmental role in human vision. There is psychological evidence that infants are sensitive to a broad range of apparent anomalies in the causal structure visual motion events (e.g., objects making contact without transferring momentum) even before they can reliably segment still objects—let alone recognize them (Spelke & Van de Valle 93).

This paper describes early results in video understanding and “gisting”—summarizing the action in

sequences. The target domain is “how-to” videos—how to put together a chair, how to open a computer, how to install a CD-ROM, etc. The input is a video of an object being assembled or disassembled. The output is a script describing the actions of the repairman plus key frames that highlight important causal events. In this paper we show a cartoon summary generated from a video; the output also supports sophisticated reasoning about the causal evolution of the scene.

2 Related work

Early approaches to action understanding emphasized reconstruction and analysis; lately attention is turning to applying causal constraints directly to motion traces. Kuniyoshi & Inoue (93) and Ikeuchi & Suehiro (94) presented substantial systems that recognize actions in assembly tasks with simple geometric objects, e.g., blocks. These early systems depended heavily on accurate reconstructions of the scene, to the degree that some actions are defined in terms of material-specific features, e.g., Kuniyoshi uses alignments of edges and changes in brightness.

Focusing purely on motion patterns, Siskind & Morris (96) demonstrated a system that distinguishes individual throwing, dropping, lifting, and pushing gestures on the basis of relative velocity profiles between an arm and an object. In a sophisticated treatment of the spatial structure of causal relations, Mann, Jepson, & Siskind (96) developed a system that analyzes kinematic and dynamic relations between objects in each frame. The system finds minimal systems of Newtonian equations that are consistent with each frame (via a preference semantics), but these are not necessarily consistent over time nor do they mark causal events.

All of these systems require both a priori knowledge of the scene (e.g, hand-segmentation of event boundaries or objects), and limited scenes (e.g., white/black backgrounds and/or constraints on the shapes and colors of objects).

In contrast, the methods described in this paper emphasize globally consistent causal structure over time and use psychologically plausible representations to qualitatively characterize local causal structure in space. This combination yields meaningful output with virtually no *a priori* knowledge about the specific scene, e.g., the

background may be cluttered and objects may be textured, irregular and flexible.

3 Segmentation of the agent

The gister reasons about changes in motions and contact relations between participants in an action. Usually only the agent (typically some part of an arm) and any objects under its control are segmented; other objects and surfaces in the scene are inferred once they participate in an action. Similarly, it suffices to extract only moving surfaces, leaving to later inferences the issues of whether those surfaces belong to one object and whether all the participants in an action have actually been segmented.

For computational ease, segmentation of moving surfaces is approximated by identifying and grouping pixels that have changed from the background. Using a precursor of the system described by Wren, Azarbayejani, Darrell, & Pentland (95), a background is “learned” by acquiring statistics for color mean and variance (in YUV space) for each pixel. In subsequent frames, pixels with large deviations from the model are tagged. The largest connected set of such pixels is output as the foreground. When an object moves or is moved, its pixels “pop out” of the background into the foreground. A complementary “hole” also pops out, but because the object is connected to the arm and the hole is not, it drops from the foreground. The system has been modified so that when altered regions such as holes cease moving, they are written back into the background.

The resulting blob typically contains the agent and any objects that it propels. We are interested in the leading edge, where contact relations are most likely to be formed. This is usually the tip of the hand, or the end of any tool the hand is holding. A variety of simple morphological properties are used to estimate the leading edge of the agent:

- *peak curvature*: The most acute convexity at several scales, often the tip of a finger or a tool.
- *forward edge*: The edge that leads the blob in its recent direction of motion.
- *most-remote edge*: The edge that lies on a line drawn through the blob’s center of mass and its entry point into the frame.
- *equidistant point*: The midpoint along the perimeter, measured from where the agent enters the frame.

The leading edge is taken to be the point along the perimeter where two or more of these estimates agree with each other and with the previous location of the leading edge.

The output of foreground segmentation is the vector $(x, y, dx, dy, a, da, p(e))$, where

- x, y = position of the leading edge;
- dx, dy = motion of the leading edge;

- a, da = area of the blob and any changes to the area in front of or behind the leading edge; and
- $p(e)$ = a normalized confidence measure indicating how close the leading edge is to the boundary of any known objects (e.g., objects that have previously been moved).

The behavior of this tuple over time is the basis for causal event detection, described in the next section.

4 Causality of motion

Spelke & Van de Valle (93) describe a series of experiments showing that infants aged 7.5 to 9.5 months are sensitive to a wide range of visual events derived from causal constraints on motion. They posit that three basic principles are active in motion understanding by late infancy: contact, cohesion, and continuity.

- The principle of *contact* equates physical connectedness with causal connectedness: “no action at a distance, and no contact without action.”
- The principle of *cohesion* equates object integrity with individuality: “no splitting, no fusing.” This guarantees that the identity of objects remains stable over time, unless a series of causal events combines two objects into one (e.g., via attachment) or splits one into two (e.g., via extraction).
- The principle of *continuity* guarantees object solidity: “no gapped trajectories, no contactless collisions.” Objects must occupy every point along their trajectory (including a connected path behind anything that occludes their motion), and no two object trajectories can occupy the same space at the same time without inducing a contact relation.



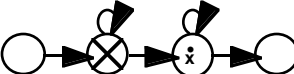









Interestingly, infants show fluent application of these constraints well before they can reliably segment still objects, leading to the speculation that aspects of appearance that remain invariant between causal events could be used to scaffold the acquisition of gestalts used to group surfaces and boundaries.

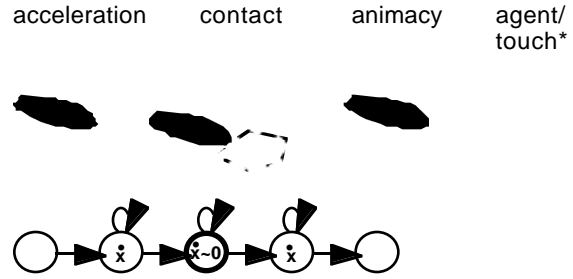
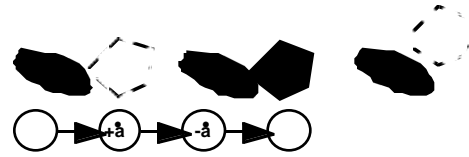
4.1 Causal events

In their experiments, Spelke and Van de Valle were able to produce apparent causal violations by using “hidden animacy” in their test stimuli, e.g., objects set in motion via invisible wires. In real life, the interaction of animate agents and inanimate objects is an enormously important part of child visual experience and learning. In order to support visual understanding in this domain, we need to add a constraint:

- The principle of *animacy* governs non-contact accelerations: “no contactless accelerations without agency or gravity.”

In our system, these four principles are used to account for the behavior of the foreground-segmented blob. When there is an apparent violation of an active constraint, this indicates that there is new information about the causal structure of the scene—usually a change in the causal connectivity, number of known individuals, or presence of the agent. The table below shows characteristic visual events, the change in governing constraints, and the meaning of the event. The figures depict a cartoon of each visual event above a transition network that detects it.

blob event	violated causality	new causality	causal event
appearance	contact	animacy	agent/enter
			
			
disappearance	contact	animacy	rest/exit
			
			
inflation (fusing)	cohesion	contact	attach
			
			
deflation	cohesion	contact	detach
			
			
flash	cohesion	contact	bump



4.2 Event detection

These networks compete to account for the input stream, calculating confidence measures on the basis of expected rates of motion (\dot{x}) and expansion (\dot{a}) and accordance with expectations from the higher-level semantics (detailed below). Bold-bordered nodes indicate likely key frames, as in the moment of contact for an attachment. They also indicate special calculations, as in the touch network, which looks for lingering near the boundary of a known object followed by a change of direction. Other networks not shown recognize simple motion.

5 Interpretation of action

Event detection by itself is neither accurate nor meaningful. We employ a higher-level action semantics to resolve ambiguity by enforcing causal constraints that have a longer temporal course, for example, that “what goes up must eventually come down.” These semantics describe sequences of events that are possible in a world that behaves sensibly.

The manipulation semantics is expressed as an action grammar:

```

scene -> in action* out
action -> motion | move | (out in)
in -> ENTER | add
out -> LEAVE | remove
add -> ENTER motion* DETACH

```

remove -> ATTACH motion* LEAVE

move -> ATTACH motion+ DETACH

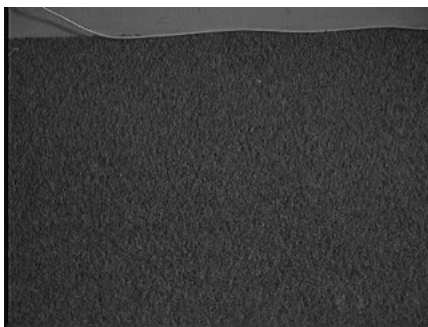
motion -> SHIFT | TOUCH | BUMP

This positive semantics implicitly encodes a number of constraints on manipulation, including: Objects cannot enter or leave the scene by themselves; an object must be put down before another can be picked up; and the arm must re-enter after exiting the scene. It may prove more desirable to express these constraints explicitly as illegal states; at present we are restricting the semantics to positive states in preparation for porting the entire system to a Markovian framework.

Video is processed by expanding the action grammar until possible causal events are reached (as the video begins, only ENTER is possible), then matching the predicted transition networks to the incoming blob descriptions. Alternative parses are kept in memory and scored for their fit to the data. The scoring function is heuristic, rewarding parses for the number of causal events that they explain but decaying as those events recede into the past. The top-ranked parses are kept from frame to frame; parses that fall below a threshold are discarded and replaced with a small number of new parses. This process continues until a high-ranked parse of an **in**, **out**, or **move** action completes. The parse is accepted, written into the script along with landmark frames, and all rival parses are discarded. Any new object boundaries revealed by the action are written into a scene list, and the system begins anew expanding from the **action** rule.

6 Example

This section details an explanation of a 500-frame sequence processed by a pilot version of the gister. The sequence shows how to open a computer housing. The gister generates an action script and picks out 7 key frames. The sequence of figures shows the output. In figure 1, for example, we show the extracted key frame, blobs of adjacent frames, the visual event (**DETACH**), the parse path that explains the event (**scene:action:in:add**), and the tag emitted by the parser (**Put**).



0: Start



1: *Put* = scene:action:in:add:DETACH



2: *Put* = scene:in:add:DETACH



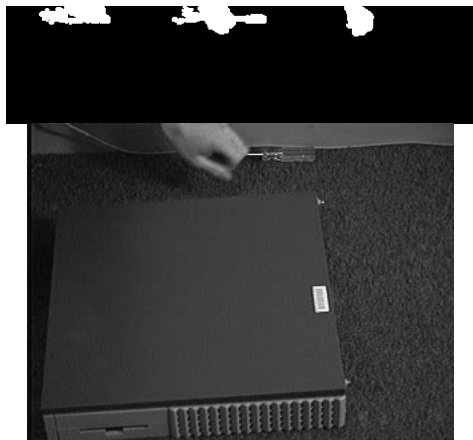
3: *Get* = scene:action:move:ATTACH



4: Touch =
scene:action:move:motion:TOUCH



5: Touch =
scene:action:move:motion:TOUCH



6: Put = scene:action:move:DETACH

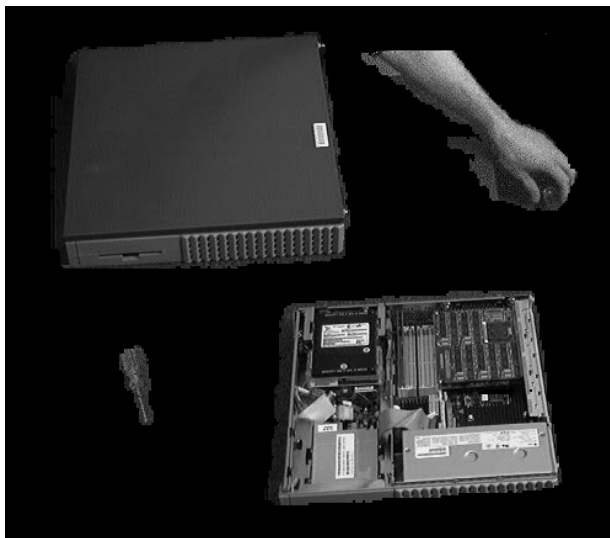


7: Remove = scene:out:remove:ATTACH



8: End frame

This cartoon summary was generated by tagging all frames that the parse resolved to ATTACH, DETACH, or TOUCH actions—actions that are interesting because they change the causal structure of the scene. Of course the parse contains (or implies) much more information than just frames and tags. It may also be used to edit the film down to a highly informative short, to log parts of the film for fast access, to index the video for retrieval from a library, or for higher-level reasoning about the structure of the scene or the intentions of the actor. For example, the object obtained in key frame 3 is not put down until key frame 6; it is in-hand during the touch events of key frames 4 and 5, indicating that it has been used as a tool. Similarly, it is possible to infer that there has been a net decrease in the number of object parts in the scene, implying a disassembly. More visually, the script and blob information make it possible to segment out the actors:



Actors in the disassembly movie

7 Limitations and future work

The gister is currently limited to sequences in which the actor is a human arm acting upon cohesive non-liquid objects (of arbitrary shape). It understands touching, putting, getting, adding, and removing actions, using evidence from crude foreground segmentations, constraints from physical causality, and context from recent actions. It also understands bumping actions, but frequently confuses these with spurious inclusions of shadows in the foreground. We are currently in the process of moving to a more sophisticated and robust blob tracker which is less sensitive to changes in illumination and also supports multiple animate agents.

The system is currently oblivious to occlusion events, e.g., the hand passing behind objects or objects being extracted from and deposited in enclosures. A semantics for these kinds of events has been written and will eventually be connected to visual routines that exploit the stability of occlusion boundaries. We are keenly interested in how new forms of causality are noticed and learned. For example, novel kinds of objects, e.g., liquids and divisible substances (clay, sand), will require more acuity, state, and substantially elaborated semantics. Not surprisingly, it takes children years to develop a perceptual and cognitive understanding of these substances.

Readers may have noticed that the blob vector and mid- and high-level semantics are exactly suited to parsing via hidden Markov models. We are in the process of porting the entire system from heuristic parsing to a Markovian framework, which will enable learning of near-optimal models for relating visual evidence to the action semantics. This will also set the stage for research efforts in learning new forms of causality and in developing hybrid temporal/spatial Markovian classifiers.

9 Conclusion

Understanding the causality of motion is a central task in vision—infants do it before they have robust vision of static objects, and adults must do it to plan interactions with the world. We present semantics and techniques for interpreting video of manipulation tasks such as disassembly, and an implementation that produces a script of the causal evolution of the scene. Psychologically-based causal constraints are used to detect meaningful changes in the integrity and motions of foreground-segmented blobs; a small causal model of manipulation is used to parse these into a coherent account of video's action. Causal landmarks such as putting, touching with a tool, removing, etc., index directly to key frames for a cartoon summary, and foreground masks can be used to segment participating objects from these frames. The output is suitable for gisting, indexing, automated video editing, and plan analysis.

10 Acknowledgements

Many thanks to Sandy Pentland and Nuria Oliver for helpful comments.

References

- [1] K. Ikeuchi and T. Suehiro (1994). Towards an assembly plan from observation, part I: Task recognition with polyhedral objects. In *IEEE Transactions on Robotics and Automation*, vol. 10, num. 3.
- [2] Y. Kuniyoshi and H. Inoue (1993). Qualitative recognition of ongoing human action sequences. In *Proceedings, IJCAI93*.
- [3] R. Mann, A. Jepson, and J.M. Siskind (1996). The computational perception of scene dynamics. In *Proceedings, ECCV-96*.
- [4] S.J.M. Siskind and Q. Morris (1996). A maximum-likelihood approach to visual event classification. In *Proceedings, ECCV-96*.
- [5] E.S. Spelke and G. Van de Valle (1993). Perceiving and reasoning about objects: Insights from infants. In N. Eilan, R. McCarthy and W. Brewer, eds. *Spatial Representation*. Oxford: Basil Blackwell.
- [6] C. Wren, A. Azarbayejani, T. Darrell and A. Pentland (1995). Pfunder: Real-time tracking of the human body In *SPIE Proceedings* vol. 2615.