The "Inverse Hollywood Problem": From video to scripts and storyboards via causal analysis

Matthew Brand
The Media Lab, MIT
20 Ames Street, Cambridge, MA 02139 USA
brand@media.mit.edu; www.media.mit.edu/~brand

Abstract

We address the problem of visually detecting causal events and fitting them together into a coherent story of the action witnessed by the camera. We show that this can be done by reasoning about the motions and collisions of surfaces, using high-level causal constraints derived from psychological studies of infant visual behavior. These constraints are naive forms of basic physical laws governing substantiality, contiguity, momentum, and acceleration. We describe two implementations. One system parses instructional videos, extracting plans of action and key frames suitable for storyboarding. Since learning will play a role in making such systems robust, we introduce a new framework for coupling hidden Markov models and demonstrate its use in a second system that segments stereo video into actions in near real-time. Rather than attempt accurate low-level vision, both systems use high-level causal analysis to integrate fast but sloppy pixel-based representations over time. The output is suitable for summary, indexing, and automated editing.

Introduction

A useful result from a vision system would be an answer to the question, "What is happening?" This is a question about causality — What are the events and how do earlier ones cause or enable later ones? We are exploring the hypothesis that causal perception rests on inference about the motions and collisions of surfaces (and proceeds independently of processes such as recognition, reconstruction, and static segmentation). In this paper we present two computational models of this process — one heuristic, one probabilistic and trainable — that incorporate psychological models of causal event perception in infants. These systems use causal landmarks to segment video into actions, and higher-level causal constraints to ensure that actions are consistent over time. Each system takes a video sequence of manipulative action as input, and outputs a plan-of-action and selected frames showing key events — the "gist" of the video — useful for summary, indexing, reasoning, and automated editing. Gisting may be thought of as the "inverse Hollywood problem" — begin with a movie, end with a script and storyboard.

Appears in Proceedings of AAAI97, Providence, RI.

Related vision work

Early approaches to action understanding emphasized reconstruction followed by analysis; lately attention is turning to applying causal constraints directly to motion traces. (Kuniyoshi & Inoue 1993) and (Ikeuchi & Suehiro 1994) described systems that recognize actions in assembly tasks with simple geometric objects, e.g., blocks. These systems were intended as front ends for robotic pick-and-place mimicry and emphasized scene geometry, taking somewhat ad hoc approaches to causality and action.

Presently there is a growing literature in gesture recognition from motions (Essa 1996), with an emphasis on classification rather than interpretation of structured activity. (Siskind & Morris 1996) blurs this distinction somewhat by using Markov models to classify short sequences of individual motions as throwing, dropping, lifting, and pushing gestures, given relative velocity profiles between an arm and an object. (Mann, Jepson, & Siskind 1996) present a system that analyzes kinematic and dynamic relations between objects on a frame-by-frame basis. The program finds minimal systems of Newtonian equations that are consistent with each frame, 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; specific camera views; and constraints on the shapes and colors of objects). In contrast, the methods described in this paper emphasize continuous action parsing, integration of information over time, constraints derived from psychological experiment, meaningful output, and general vision, e.g., the background may be cluttered and objects may be textured, irregular and flexible.

Psychology of motion causality

Vision sciences traditionally take high-level vision to be concerned with static properties of objects, typically their identities, categories, and shapes. The relationships between these properties and visual features are *correlational*, leading to many proposals for how brains and computers may compute optimal discriminators for various sets of images.

Arguably, causal dynamic properties of objects and scenes are more informative, more universal and more easily computed. These properties — substantiality, solidity, contiguity, inertia, and conservation of momentum — are governed by simple physical laws at human scales and are thus consistent across most of visual experience. The fact that these properties are *causal* suggests that a small number of qualitative rules may provide satisfactory psychological and computational accounts of much of visual understanding.

Indeed, there is a growing body of psychological evidence showing that infants are fluent perceivers of lawful causality and violations thereof. Spelke and Van de Valle found that infants aged 7.5 to 9.5 months will detect a wide range of apparent violations of the causality of motion (Spelke & Van de Valle 1993). They propose that three basic principles are active in motion understanding by late infancy:

- The principle of **contact** equates physical connectedness with causal connectedness: *"No action at a distance; no contact without action."*
- The principle of **cohesion** equates object integrity with individuality: *"No splitting, no fusing."* This guarantees that individuality (boundaries) remain 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 detachment).
- 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 occluders), and no two objects can move into the same space without inducing a contact relation.

Note that that these conservation laws are equally applicable to collections of objects *or* surfaces. This suggests that it may be possible to discover the event structure of a video by reasoning about the motions and collisions of surfaces. With some amendments, this is the strategy used by two vision systems described below.

We close with one remarkable finding: Infants often achieve fluent application of causal constraints to motion but still make systematic errors in static object segmentation. We may speculate that mastery of motion causality has a developmental role in general vision: Aspects of appearance that remain invariant *between* causal events could be used to scaffold the acquisition of surface- and boundary-grouping gestalts.

Heuristic gisting of "how-to" videos

We turn now to the problem of instantiating causal constraints in a computer vision system. Since manipulatory action is rich in causal structure, we sought to develop a "video gister" — a system that could extract key events from "how-to" videos of the sort that demonstrate procedures for assembling furniture, installing CD-ROMs, 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.

From visual events to causal events

The gister reasons about changes in the integrity and motions of a single foreground blob — a connected map of image pixels that change due primarily to motion. The blob is obtained from a real-time vision system developed by (Wren *et al.* 1995). Discontinuities in the blob's visual behavior signal changes of causality. For example, if the blob has a boundary discontinuity such as sudden swelling at one point, there is an apparent violation of the **cohesion** constraint, explicable via the **contact** constraint: An agent has attached an object and set it in motion, causing its pixels to join the blob. (Cohesion is violated because the agent "fuses" with the object.) Many visual discontinuity events have causal significance, including:

visual event	disrupted causality	explanatory causality	causal event
appearance	contact	animacy	enter
disappearance	contact	animacy	leave
inflation	cohesion	contact	attach
deflation	cohesion	contact	detach
flash	cohesion	contact	bump
acceleration	contact	animacy	touch
discontinuity	continuity	cohesion	occlusion

(A flash is a sudden appearance and disappearance of a motion surface.) Note that we have introduced a fourth fundamental constraint:

- The principle of **animacy** governs non-contact accelerations: *"No contactless accelerations without agency or gravity."*

When a causal event is detected, the causal explanation imparts some new information about the scene. For example, when a blob appears (due to spontaneous motion), the unexplained acceleration of a surface violates the **contact** constraint, but is explained by the **animacy** constraint, imparting the information that the agent has been located. Similarly, an attachment is explained by the **contact** constraint, and imparts the information that an object has been removed from the area co-located with the blob growth and is now "in hand." The gister accumulates this information over time to build a rough scene model.

Event detection alone is not accurate enough to yield a meaningful analysis. We must take recourse to higher-level causal constraints which enforce longer-term consistencies in the video parse. The high-level constraints, in this case a semantics of manipulation, are expressed as an action grammar. It adds to the basic causal laws some 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 agent cannot leave twice without reentering:

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

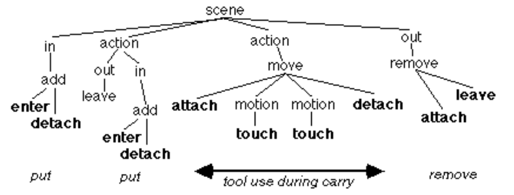

Figure 2: Parse tree of the disassembly video.

Implementation

Analysis begins with the foreground blob, which 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. An ensemble of morphological operators are used to estimate the leading edge of the agent, including *peak curvature* (multiscale estimate of the contour's sharpest point); *forward edge* (motion-leading edge); *remote edge* (point most remote from the blob's entry into the frame); and *equidistant point* (perimeter point equidistant from blob's entry into frame). The leading edge is taken to be the point of greatest agreement between two or more of estimates.

Once the leading edge is found, the blob is characerized by the vector $(x, y, \dot{x}, \dot{y}, a, \dot{a}, p(e))$, where x, y locate the leading edge; $\dot{x}, \dot{y}$ track its velocity; $a, \dot{a}$ measure the area of the blob and any changes to the area in front of or behind the leading edge; and $p(e)$ is 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 vector over time is the basis for causal analysis.

Visual events are detected via small finite-state transducers (FSTs) with Gaussian output parameters. These are akin to left-right hidden Markov models, except that instead of transition probabilities, each state is assigned a mean duration. States may also emit tags, e.g., if the FST for deflation is accepted, it emits a PUT tag.

The FSTs compete to account for segments of the vector stream; a high-scoring FST indicates a likely visual event in its time span. An FST is fitted to a set of blob vectors by adding random values to the durations, calculating the probabilities of each vector given the corresponding state's (Gaussian) output parameters, then taking the geometric average of the result as a score.

A stream of blob vectors is processed by expanding the action grammar until possible causal events are reached, then matching the predicted transition networks to the incoming blob descriptions. Approximately 100 competing 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 threshhold are discarded. This process continues until a high-ranked parse of an in, out, or move action completes. The parse is accepted and written into the script along with landmark frame tags that are emitted by special nodes in the FSTs. All rival parses are discarded. Any new surface boundaries revealed by the action are written into a scene list, and the system begins anew expanding from the action rule.

Example

Here we review the results of processing a 500-frame video sequence showing how to open a computer housing. The video gister generated a 13-production action parse and picked out 7 key frames associated with causal events of interest (ATTACH, DETACH, TOUCH). The key frames, captioned with their associated event tags, are montaged into the storyboard shown in figure 1. The parse tree is shown in figure 2. The parse contains some very useful information: It might 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 objects in the scene, implying a disassembly. More visually, the script and blob information make it possible to segment out the actors. Once an object has been moved, the "hole" in the background field gives us its boundaries. The script allows us to find an earlier frame in which the object is visible and undisturbed; the subsequent hole is used as a template for cutting the object out.

Limitations

This 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,

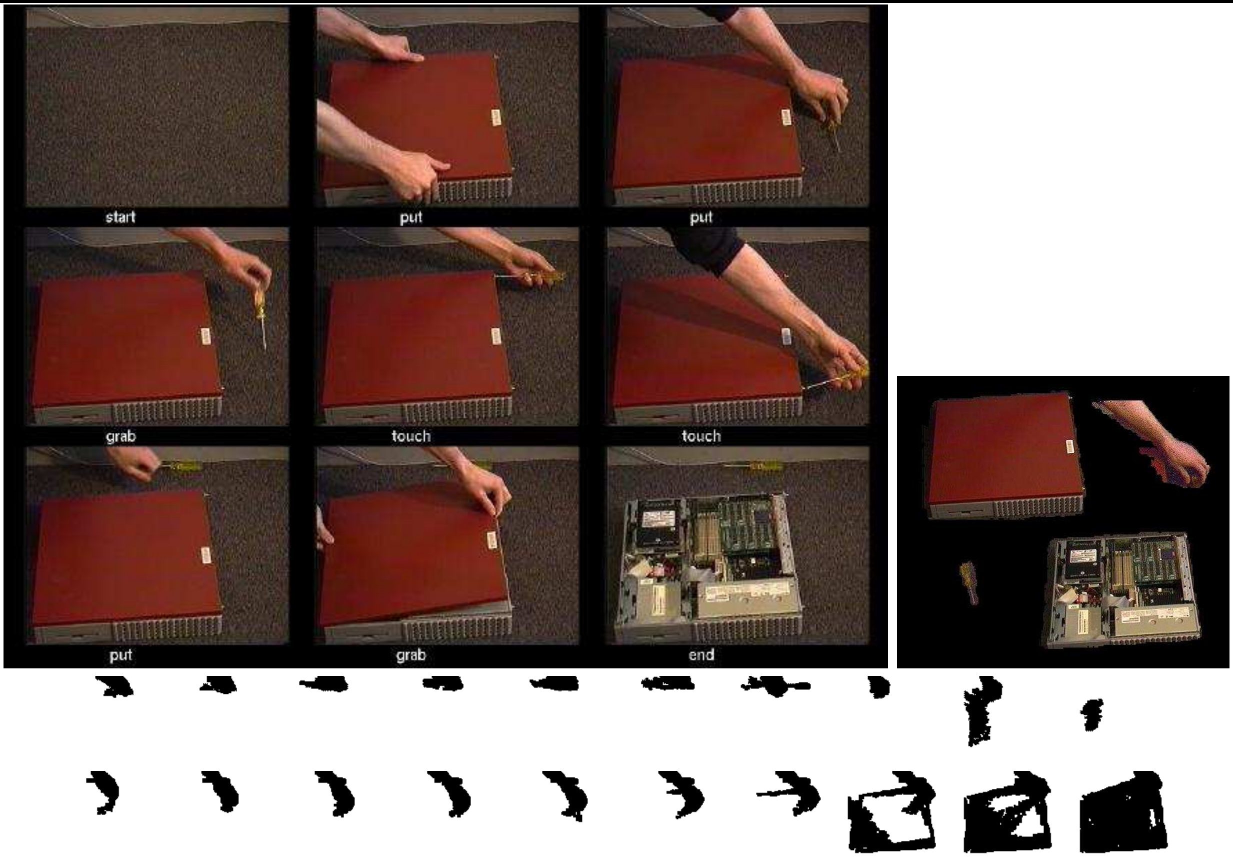

Figure 1: Storyboard of the disassembly video. The tags are start, put, put, grab, touch, touch, put, grab, end. On the right, the "actors" have been segmented. Underneath are blob sequences of the final put and grab.

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. The event detectors and parse scoring algorithm are hand-tuned and therefore not adaptive or robust. The blob-tracker will only register a single connected moving mass, effectively blinding the system to throwing, batting, and other actions that impart momentum to inanimate objects. Finally, the system runs approximately 1/10th real time. Our second system specifically addresses these challenges, tracking multiple moving objects and learning probabilistic event detectors in a expectation-maximization framework.

Probabilistic gisting of live action

We turn now to formulating causal motion analysis as a learnable probabilistic inference task. If the the causality of the world is expressible as a finite-depth rule system and the event detectors are parameterized probability models, the two can be combined into a global probability model, opening the way to maximum-likelihood interpretation of entire action sequences. Here we follow in the spirit of continuous speech recognition systems, showing how to compile action-parsing systems into large stochastic models.

We raise the stakes, however, by taking on the problem of recognizing actions involving three parties, such as batting, which requires an agent (batter), an instrument (bat), and a recipient (ball). Many examples can be found in three-place pragmatic verbs in language ("X gave Y the Z; U placed the V on the W;" etc.). The strategy used in (Siskind & Morris 1996) — reducing two-party actions to a single spatial relation which is then tracked by a hidden Markov model (HMM) — will not work because there are a variety of relations between three objects whose relative importance varies over time. A system that has several important variables whose variation cannot be mutually predicted is said to have compositional state. Unfortunately, Markov models deal very poorly with such systems, because the probability model is founded on the assumption that the system can be efficiently described by a single discrete state variable.

Coupled hidden Markov models

To address the problem of compositional state, we introduce a method for coupling and training hidden Markov models. We sketch the method here; a full mathematical exposition can be found in (Brand 1996). An HMM is a quantization of a system's configuration space into discrete regions $S = \{a_1, a_2, a_3, \ldots, a_N\}$, which are selected by a time-indexed state variable $s_t \in S$. A transition matrix $P(s_t = a_i | s_{t-1} = a_j)$ describes the probability that the system is in a region given its previous location. A second set of parameters describes the probability of seeing an observation given each hidden state. Because the state variable cannot be directly inferred from an observation, we say that it is hidden. There are efficient $O(TN^2)$ algorithms to estimate most likely states sequences and parameters given observations sequences.

We define a distribution over K individual HMMs by adding coupling matrices $P(s_t^{(a)} = a_i | s_{t-1}^{(b)} = b_j)$, which are analogous to transition matrices except that they condition each HMM state on the previous states of *other* HMMs. The increased degrees of freedom make precise state and parameter estimation $O(TN^{2K})$, which is typically infeasible due to data and time scarcities. We have introduced an efficient algorithm which minimizes cross-entropy with the true distribution, yielding high quality approximations in $O(T(KN)^2)$ or $O(TKN^2)$, depending on the degree of coupling. Coupled HMMs have substantially outperformed conventional HMMs in a number of simulated and real vision tasks, with better data fitting, classification, and robustness to initial conditions (Brand & Oliver 1996).

Implementation

Using coupled HMMs as our event detectors, we assembled a vision system as follows:

The motion blob tracker was replaced with two vision systems that each track three blobs of connected pixels of similar color, regardless of motion. These two systems are arranged in an uncalibrated wide-disparity stereo vision rig, and feed their results to a third system which recursively estimates maximum-likelihood 3D positions of the objects (Azarbayejani & Pentland 1996). To remove high-frequency noise, the stream of 3D positions is resampled at 60Hz, low-pass filtered, and downsampled. The result is then differenced both spatially and temporally, yielding object velocities $\dot{x}$, relative distances d, and relative velocities $\dot{d}$.

The four events we sought to detect were: picking up an object; putting it down; pushing an object; and batting one object with another. We obtained six samples of each action for training; examples are shown in figure 3. We reduced the three-party interactions to two spatial relations, a strategy licensed by our ability to identify the agent as the blob that moves first. (We note an interesting coincidence: (1) Four-party actions are quadratically more difficult to learn in the coupled HMM framework. (2) There are no four-party pragmatic verbs in any natural language.) Coupled HMMs were trained using the dual vectors $(\dot{x}_a, \dot{x}_1, d_{a1}, \dot{d}_{a1}, d_{12}); (\dot{x}_a, \dot{x}_2, d_{a2}, \dot{d}_{a2}, d_{12})$, where $\dot{x}_a$ is the velocity of the agent, $\dot{x}_1$ the velocity of object one, and d_{a1}, d_{a2}, d_{12} are distances between the agent and objects one and two. The training data was permuted with the objects swapped so that the system would learn that all inanimate objects have the same causalty. For comparison, conventional HMMs were also trained using the full state vector.

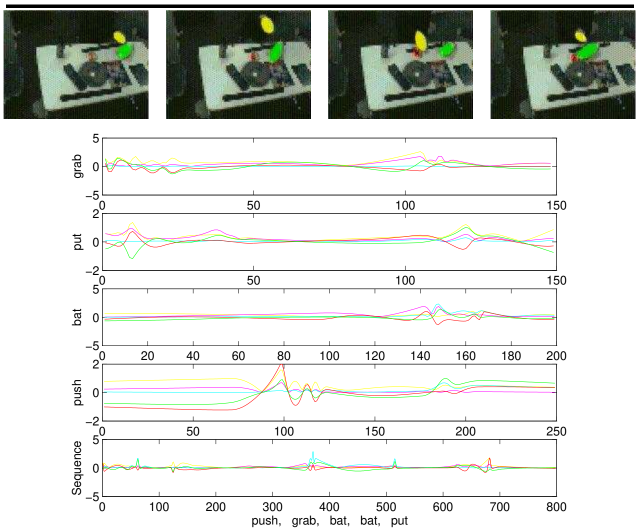

Figure 3: *Images*: Snapshots of grab, put, push, and bat actions. *Top four plots*: Their velocity profiles. *Bottom plot*: A short sequence of continuous action.

The resulting HMMs were then linked into a super-HMM according to the finite-depth grammar

$$
\begin{aligned}
\textsf{scene} &\rightarrow \textsf{act}+ \\
\textsf{act} &\rightarrow \textsf{push} \mid \textsf{move} \mid \textsf{tool} \\
\textsf{tool} &\rightarrow \textsf{grab}\ \{\textsf{carry} \mid \textsf{bat}\}^*\ \textsf{put}
\end{aligned}
$$

This encodes the commonsense constraints that picking up and putting down must alternate, and that one needs a tool for batting. (Carry and move are one-state HMMs with large output variances that accommodate free uneventful motion.) This grammar is analogous to the previous grammar, but lacks rules governing entering, exiting, and touching previously detected surfaces because the Markov formulation does not accommodate an accumulated scene model. For comparison, HMMs were also linked without causal rules, essentially allowing arbitrary sequences of events.

Results

For testing, the training materials were discarded and replaced with a new set of objects. Eight action-packed sequences of 500-2000 frames were captured and the actions performed were hand-logged in a ground truth file. The four super-HMM variants (±coupling ±grammar) were used to segment and label the sequences by computing the maximum likelihood state path through their component event HMMs. The resulting event sequences were aligned with the ground truth and statistics were collected for true hits (H=# correct labels) false alarms (F=#excess events) and false rejections (M=#missed events). From these we calculated recall $R = H/(H + M)$ and precision $P = H/(H + F)$ scores for the four actions:

model	H	M	F	R	P
HMM	20	17	15	54%	57%
+grammar	28	9	9	75%	75%
+coupling	30	7	9	81%	76%
+coupling+grammar	36	1	2	**97%**	**94%**

The combination of a better probability model (coupled HMMs) and causal constraints (grammar) produced far superior results, with nearly perfect labelings (the miss and false alarms are all pushes, which have very faint motion signatures). As we remove these advantages, performance falls off rapidly. The data suggest that the grammar principally reduces false alarms, while the coupling reduces the number of misses by improving detection.

Limitations and future work

The architecture of the probabilistic video gister is trainable, robust, capable of three-party actions, and near real-time. In these regards it improves enormously over the first system. The consistent probabilistic framework will facilitate moving to more robust surface trackers and may even provide priors for probabilistic low-level vision algorithms. On the other hand, a purely probabilistic framework is not (yet) hospitable to representational feats such as incrementally building and consulting a scene model, a matter for future research.

We are keenly interested in how forms of causality are noticed and learned, and their relation to language. An intriguing and mathematizable hypothesis is that causal laws may be discovered as variants of a more fundamental conservation law that divides visuotemporal experience into phases in which visual properties change slowly, separated by sharp regime transitions of causal import.

Conclusion

We address the problem of visually detecting causal events and fitting them together into a coherent story of the action witnessed by the camera. We show that this can be done by reasoning about the motions and collisions of surfaces. Motion causality has a clear semantics that has largely been worked out in psychological studies of infant visual behavior. We instantiate these constraints in two computer vision systems that produce meaningful text descriptions of video. Our first system heuristically parses instructional repair videos into action scripts and storyboards of key frames. It sees one- and two-participant actions (grab, touch, etc.) and can infer some three-participant actions (e.g., tool-use) from context. We developed a second, trainable system that perceives one-, two-, and three-participant actions using a recently developed technique for training couplings of hidden Markov models. Using these models, we compiled a causal semantics into a trainable maximum-likelihood continuous-action parser. We tested this on live stereo video of people performing complex action sequences and achieved 94% accurate action summaries. The higher-order models and the causal semantics both conferred significant advantages over standard hidden Markov models, which performed only slightly better than 50%. The success of the probabilistic method in particular suggests that causality may indeed be a meaningful vision result that computers can obtain reliably and efficiently.

References

Azarbayejani, A., and Pentland, A. 1996. Real-time self-calibrating stereo person tracking using 3-D shape estimation from blob features. In *Proc. ICPR96*.

Brand, M., and Oliver, N. 1996. Coupled hidden markov models for complex action recognition. In *Proc. CVPR97*.

Brand, M. 1996. Coupled hidden markov models for modeling interacting processes. *Forthcoming (under review)*. Available as MIT Media Lab Vision & Modeling TR #405.

Essa, I., ed. 1996. *Proc., 2nd International Conf. on Automatic Face and Gesture Recognition*. Killington, VT: IEEE Computer Society Press.

Ikeuchi, K., and Suehiro, T. 1994. Towards an assembly plan from observation, part 1: Task recognition with polyhedral objects. *IEEE Transactions on Robotics and Automation* 10(3).

Kuniyoshi, Y., and Inoue, H. 1993. Qualitative recognition of ongoing human action sequences. In *Proc. IJCAI93*.

Mann, R.; Jepson, A.; and Siskind, J. M. 1996. Computational perception of scene dynamics. In *Proc. ECCV96*, II:528–539.

Siskind, J., and Morris, Q. 1996. A maximum-likelihood approach to visual event classification. In *Proc. ECCV96*, II:347–360.

Spelke, E. S., and Van de Valle, G. 1993. Perceiving and reasoning about objects: Insights from infants. In Eilan, N.; McCarthy, R.; and Brewer, W., eds., *Spatial Representation*. Oxford: Basil Blackwell.

Wren, C.; Azarbayejani, A.; Darrell, T.; and Pentland, A. 1995. Pfinder: Real-time tracking of the human body. In *SPIE Proceedings*, volume 2615.