Using Causal Induction in Humans to Learn and Infer Causality from Video

Amy Fire (amy.fire@ucla.edu) and Song-Chun Zhu
Center for Vision, Cognition, Learning and Art, University of California, Los Angeles

Introduction

Goal: Computational model for the learning of causality from raw video

Motivation: Model inference processes

1. Answer why events occur
2. Correct misdetections and infer hidden/ambiguous objects/actions
3. Infer triggers, goals, and intents

Perceptual Causality

Infants use heuristics in judging causal relationships:

1. Agentive actions are causes
2. Measure co-occurrence between action A_i and effect ΔF_j

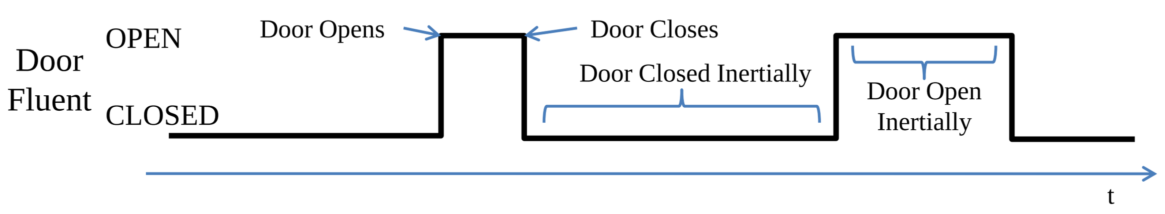
cr :		ΔF_j Present	ΔF_j Absent
	A_i Present	f_0	f_1
	A_i Absent	f_2	f_3

3. Temporal lag between the two is short
$$\text{Time(Action)} - \text{Time(Effect)} < \epsilon$$
4. Cause precedes effect
$$\text{Time(Action)} - \text{Time(Effect)} > 0$$

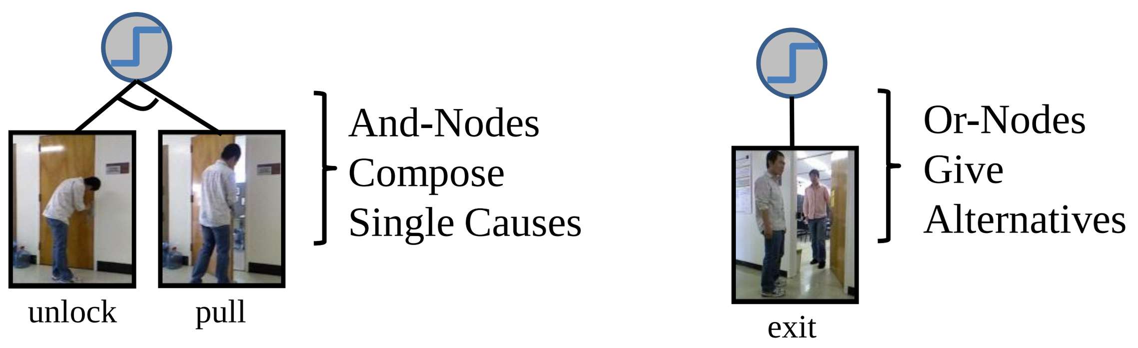
To learn perceptual causality in video, we restrict co-occurrence of detected events and effects to these heuristics.

A Grammar Model for Causality

Effects: Fluents are time-varying statuses of objects.



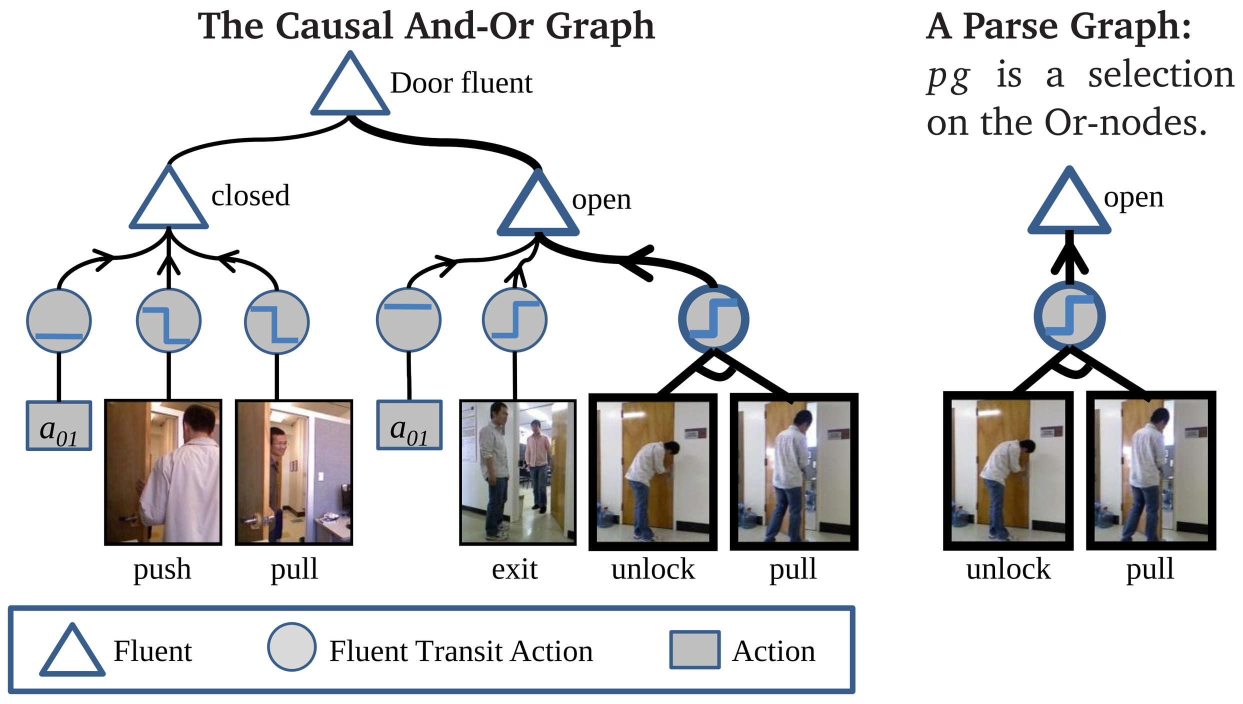
Causes: Actions suggest an And-Or representation.



Pairing cause and effect: Fluent changes are matched with corresponding causing actions. In the absence of change-inducing actions, fluent values are causally attributed to the *inertial action*, a_{01}.

The Causal And-Or Graph

A Parse Graph: pg is a selection on the Or-nodes.

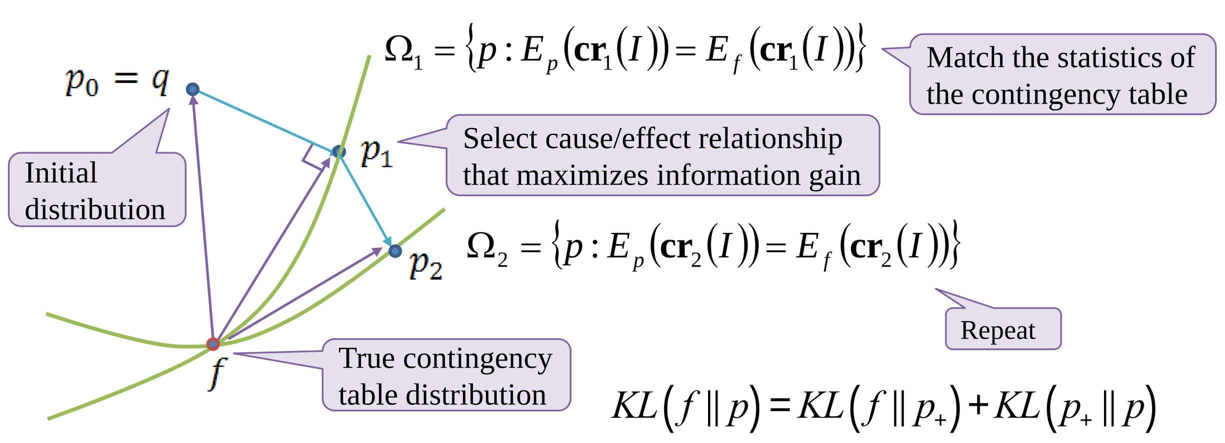


Probability on the C-AOG: Given the video I,

$$\underbrace{P(pg_C|I)}_{\text{posterior}} = \underbrace{P(A_1,\dots A_n|I)P(\Delta F_1,\dots,\Delta F_m|I)}_{\text{likelihood}} \prod_{v\in V_C^{Or}} \underbrace{P(w(v))}_{\text{prior}}$$

- likelihood: the detection probabilities
- V_C^{Or}: the set of included Or-nodes in the causal explanation
- $w(v)$: the selected Or-branch
- prior: the switch probability on the Or-nodes

Learning the C-AOG by model pursuit: Incrementally pursue a model, adding a contingency table at each iteration by information projection.



Proposition: Add the best action-fluent pair $(A_i, \Delta F_j)$:

$$\mathbf{cr}^* = \operatorname*{argmax}_{\mathbf{cr}}(\text{Information Gain}) = \operatorname*{argmax}_{\mathbf{cr}}(\text{KL}(\mathbf{f}\|\mathbf{h})),$$

where $\mathbf{f}$ is the observed frequencies of $\mathbf{cr}$ and $\mathbf{h}$ is the expected contingency table predicted by the model p in the current iteration

Experiment 1: Learning Causality

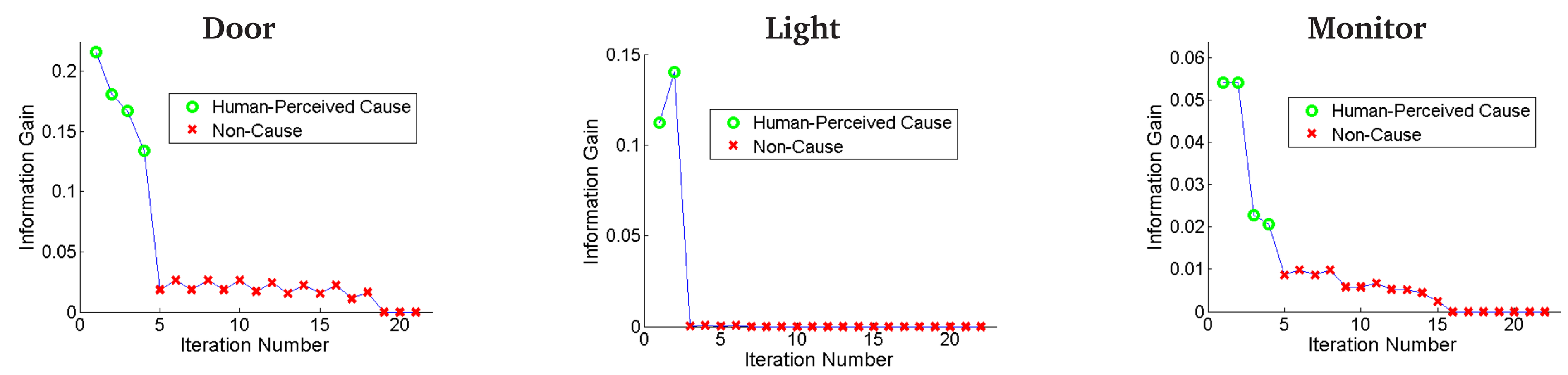
Goal: Learn causal relationships between fluent changes and actions

Methods:

- 120 minutes of video in office and hallway scenes
- 21 action categories, 8-20 instances of each
- Perfect action/fluent detection demonstrates learning
- Ground truth links known causing actions to their fluent effects

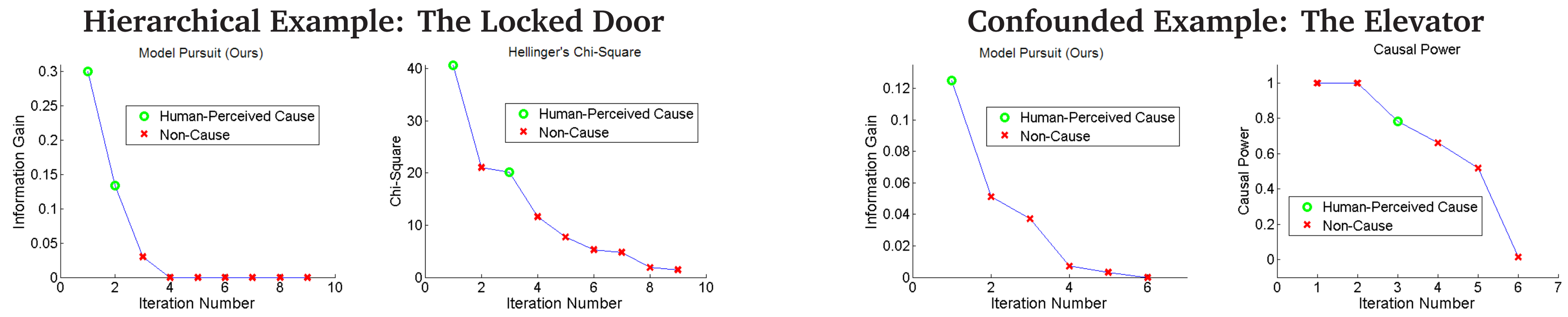
Results: Correctly matching causal relations

Information gains for causal relations in the order pursued, separated by fluent. Green circles label causes.



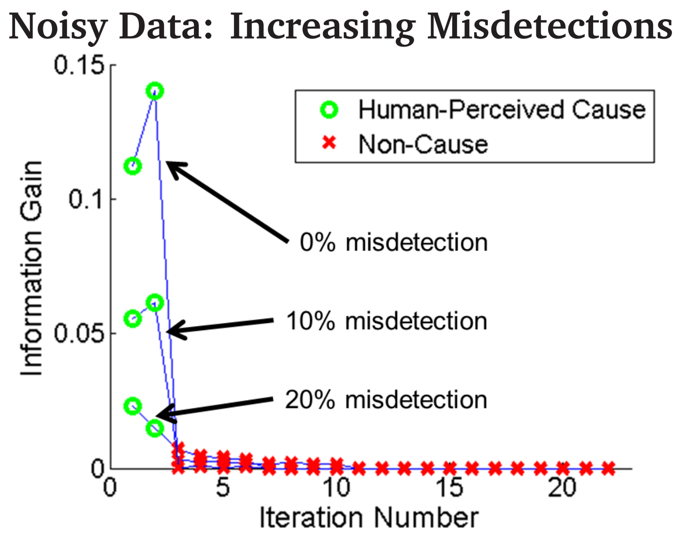
Comparisons:

Hierarchical Example: The Locked Door **Confounded Example: The Elevator**



Our method acquires true causes before non-causes, outperforming Hellinger's Chi-Square and Causal Power.

Noisy Data: Increasing Misdetections



Discussion:

- Our method matches human perceptions in the presence of multiple confusing events.
- In the presence of confounders (the monitor), our method appropriately reduces clarity in the causal relationships.
- Our method incorporates dependencies in action hierarchies (the locked door).
- Our method places importance on quantity of hits (the elevator), accommodating the ambiguity important to vision.
- Clean detections are important to being able to learn causality.
- Limitation: Our methods are limited to pre-specified action and fluent categories so that appropriate detectors can be trained.

Experiment 2: Inference Experiment

Goal: Validate our model in the long-term reasoning task of inferring hidden fluent values

Stimuli:

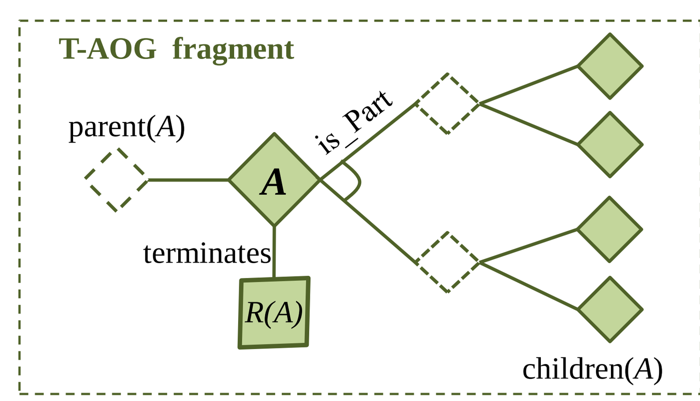
- 20 minutes of hallway and office video
- 15 volunteer participants were shown the test video which paused at preset frames surrounding fluent changes or causing actions
- Fluents shown are either ambiguous or completely hidden

List of fluents

- Computer: ASLEEP/AWAKE
- Monitor Display: ON/OFF
- Monitor Power: ON/OFF
- Cup: MORE/LESS/SAME
- Water Stream: ON/OFF
- Light: ON/OFF
- Phone: ACTIVE/STANDBY
- Trash Can: MORE/LESS/SAME
- Agent : THIRSTY/SATIATED
- Agent: HAS_TRASH/NOT

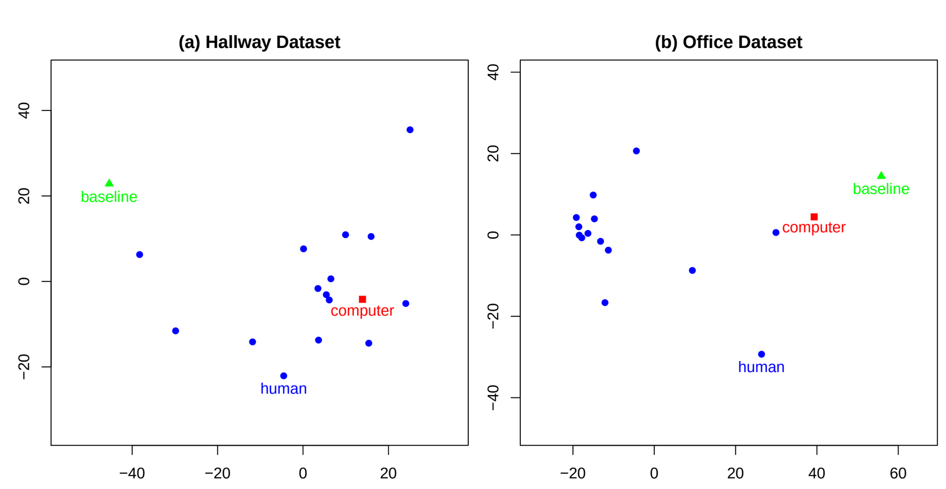
Reference estimates:

- Baseline: 50/50
- Computer (our method): From video, actions are parsed using the Temporal And-Or Graph (right) and fluent changes are extracted using GentleBoost (below). These outputs are parsed with the Causal And-Or Graph.

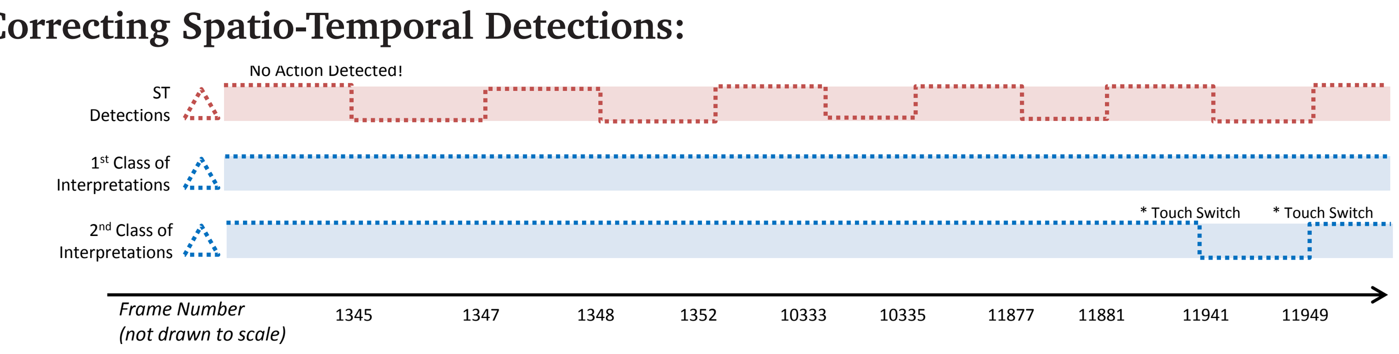


Results:

- MDS plots of fluent value estimates.



Correcting Spatio-Temporal Detections:



Discussion:

- The Causal And-Or Graph smooths over misdetections in a way that is consistent with human responses
- The Causal And-Or Graph outperforms baseline
- Variation in human responses occurs due to different initializations and different variability thresholds