Learning Perceptual Causality from Video

AMY FIRE and SONG-CHUN ZHU, University of California, Los Angeles

Perceptual causality is the perception of causal relationships from observation. Humans, even as infants, form such models from observation of the world around them [Saxe and Carey 2006]. For a deeper understanding, the computer must make similar models through the analogous form of observation: video. In this paper, we provide a framework for the unsupervised learning of this perceptual causal structure from video. Our method takes action and object status detections as input and uses heuristics suggested by cognitive science research to produce the causal links perceived between them. We greedily modify an initial distribution featuring independence between potential causes and effects by adding dependencies that maximize information gain. We compile the learned causal relationships into a Causal And-Or Graph, a probabilistic and-or representation of causality that adds a prior to causality. Validated against human perception, experiments show that our method correctly learns causal relations, attributing status changes of objects to causing actions amid irrelevant actions. Our method outperforms Hellinger's χ^2-statistic by considering hierarchical action selection, and outperforms the treatment effect by discounting coincidental relationships.

CCS Concepts: •**Computing methodologies** → **Reasoning about belief and knowledge; Computer vision; Hierarchical representations; Unsupervised learning;**

Additional Key Words and Phrases: Perceptual causality, causal induction, information projection

ACM Reference Format:
Amy Fire and Song-Chun Zhu, 2015. Learning Perceptual Causality from Video. *ACM Trans. Intell. Syst. Technol.* V, N, Article XXXX (August 2015), 22 pages.
DOI: 0000001.0000001

1. INTRODUCTION

Agents navigate the world with a perception of causes and effects. They have a deep-rooted expectation, for example, that hitting a light switch will turn the light on. Humans are equipped with the ability to form these relationships from infancy [Saxe and Carey 2006], and cognitive scientists believe that this knowledge is acquired by observation [Griffiths and Tenenbaum 2005]. Understanding this perception leads toward understanding an agent's intent and predicting his actions. Further, modeling his perceived causality connects objects and events, which can greatly improve the quality of detections amid occlusion and misdetection.

The primary analog of observation for the computer comes from video. In this paper, we propose a framework for the unsupervised learning of perceptual causal structure (specifically those causal relationships that are perceived) from video taken in everyday scenes such as an office, a doorway, and an elevator. We search here for causes of fluent changes. A *fluent* is defined in the commonsense-reasoning literature as an object status that specifically varies over time [Mueller 2006]. A door's fluent, for example, takes the values "open" and "shut" over time as the door is moved.

This work is supported by the Office of Naval Research, under MURI grant N00014-10-1-0933.
Author's addresses: A. Fire and S.-C. Zhu, Department of Statistics, University of California, Los Angeles.

DOI: 0000001.0000001

Traditional causal discovery methods are insufficient for computer vision tasks. Most importantly, true causal discovery does not necessarily align with human perceptions. Secondly, the traditional methods leave many questions: they do not inform which detection variables humans would indicate as causes or effects (from pixels over time, to features aggregating pixels, to object classifiers using features, to action detectors, to hierarchies thereof); they do not indicate how to divide the video stream to create examples (too small a clip might omit causes; too large of one introduces noise); and they do not encode a prior understanding of likely causes that could be used in detections.

Perceptual causality as studied by cognitive science researchers fills in these gaps.

Humans link, for example, a change in an object status with the action of an agent [Saxe et al. 2005]. Humans award the "cause" distinction to the agent's action of opening the door (decomposed at a high level into unlocking and pulling open the door), ahead of individual pixels, the door, and the lock (part of the door). We limit ourselves to agentive actions as potential causes of fluent changes. In order to make detections from video, these sets of actions and fluents must be pre-specified so appropriate detectors can be trained.

Considering actions alone is not enough. Actions come hierarchically defined, where, for example, the person opening the door performs the actions unlock and pull. The method we present can correctly select from a hierarchy, as shown in Section 5.3.

Humans consider cause and effect relationships when the temporal lag between the two is short and cause precedes effect [Hagmayer and Waldmann 2002]. We construct examples from the video that only consider actions occurring within a small window preceding a given effect.

Finally, humans link states of the world to perceived causing conditions by measuring co-occurrence [Griffiths and Tenenbaum 2005]. We propose a method to learn this perceptual causality, acquiring the knowledge of causal relationships in the world, illustrated by the dashed arrows in Figure 1.

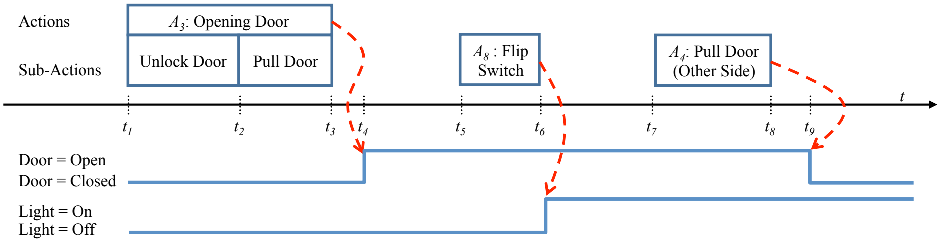


Fig. 1. Key actions over time are shown at the top, and status changes below. Dashed arrows indicate perceived causal links. A link can be found between any action in an action hierarchy and its relevant fluent change. The joint actions of unlocking and pulling cause the door to open at t_4. From t_5 to t_6, a person turns a light on. From t_7 to t_8, a person pulls the door from the other side, resulting in a closed door at t_9.

As input, our method takes potentially noisy or erratic action and fluent detections from video. Perceived causal links are then extracted from them. Our preliminary work has been published in [Fire and Zhu 2013a] and [Fire and Zhu 2013b], and is extended here.

New to this paper, we write perceptual causality characteristics as assumptions in terms of structural equation models. We then develop the theory to sequentially learn which actions cause which fluents to change. We use an information projection pursuit to learn perceptual causal structure in an unsupervised way, proving theory asserted in [Fire and Zhu 2013b]. We derive analytic solutions for obtaining and fitting the causal relations, and we show that this method selects the most relevant action from an action hierarchy.

In Section 6, the learned causal links are assembled into a Causal And-Or Graph and the learned probability model is used to encode prior information on causality.

In new experiments, we study the effects of temporal lag, finding it is best to control both the number of actions considered and the temporal lag. We further explore the number of training examples needed.

We review results from [Fire and Zhu 2013b], where we compared our results against Hellinger's χ^2 and the treatment effect, finding that our method performed best. Finally, we examine other factors that affect the perception of causality from video: incorrect detections and confounding actions.

The rest of this paper is structured as follows. After a brief summary of previous works in Section 2, we set up the necessary parts to solve our problem in Section 3. We develop our learning theory in Sections 4-5. In Section 6, we assemble the pursued causal relations into a Causal And-Or Graph. In Section 7, experiments validate the learning framework.

2. RESEARCH IN VISION AND CAUSALITY

Researchers in computer vision focus on detecting objects, understanding scenes, and analyzing actions and events. There has been growing interest in exploring contextual information to improve recognition performances. For example, image parsing work explores spatial context between objects and background [Tu et al. 2005], and video parsing work exploits temporal context between actions [Pei et al. 2011].

Disjoint from causal discovery, vision researchers have used causal context for action recognition (e.g., [Albanese et al. 2010]) and have used Newtonian mechanics to distinguish actions [Mann et al. 1997].

Separated from detection, commonsense reasoning on its own is usually solved by first-order logic [Mueller 2006]. This disallows the probabilistic solutions needed in computer vision for the ambiguity of unreliable detections. Markov logic networks [Richardson and Domingos 2006] relax the strictness of first-order logic by wrapping them in a Markov random field, and have been applied to the task of action detection [Tran and Davis 2008], but the knowledge base is not learned.

Vision researchers have used causal measures such as Granger causality to learn patterns of repeated low-level actions [Prabhakar et al. 2010]. But these methods are far from learning causal structure under traditional causal induction methods as done by constraint-based algorithms such as IC [Pearl 2009], PC, and FCI [Spirtes et al. 2000], or by Bayesian formulations that place a prior on graph structure [Heckerman 1995]. The former does not represent perceptual causality, and while the latter has been used in cognitive science [Griffiths and Tenenbaum 2005], it has not been grounded on action detections from video.

Advancing in the direction of cognitive science and perceptual causality, Brand borrows from infants' perceived implications of motion to provide the "gist" of a video using detected blobs [Brand 1997]. One of the main drawbacks to this work, however, is that the grammar is not learned.

None of these approaches formally study cause-and-effect relationships in a way that allows causal structure to be learned from video. Our method for learning perceptual causal structure, however, integrates with both spatial and temporal learning strategies. While perceptual causality lacks the accuracy of traditional causal induction, it provides valuable—and more human—information.

3. SETTING UP OUR PROBLEM

3.1. Vision and Causality: Converting Perceptual Causality to Heuristics

Perceptual causality as presented in the introduction grounds causal discovery on video, and distinguishes perceptual causality from the causal modeling typically done in the social and biologic sciences [Pearl 2009], [Rubin 2005]. We now present the heuristics of perceptual causality.

The Heuristics:

(1) Agentive actions are causes ,

$$\text{Action} \rightarrow \text{Effect}.$$

This heuristic informs the set of potential causes: It's not the pixels we see or the human that we detect, but it's the human *doing something*.

(2) Temporal lag between cause and effect is short, with cause preceding effect,

$$0 < \text{Time(Effect)} - \text{Time(Causing Action)} < \epsilon.$$

This provides a method for breaking the video stream into clips to create examples. Determining ϵ is challenging: taking it too small might exclude the cause, and taking it too large creates too much noise. We examine various temporal lags, as well as different ways of measuring the temporal lag, in Experiment 7.2.4.

(3) Perceptual causal relationships are obtained by measuring co-occurrence between actions and effects.

In this paper, we examine co-occurrence while simultaneously building our model following an information projection pursuit. In Experiment 7.3.1, we find our method outperforms Hellinger's χ^2 measure for co-occurrence. In Experiment 7.3.2, we show that our method outperforms the treatment effect.

When the computer examines the co-occurrence of Heuristic 3, restricted by Heuristics 1 and 2, then we assume the model determined represents perceptual causality.

3.2. Assumptions and Structural Equation Models

In addition to assuming Heuristics 1-3, we also make some assumptions standard to traditional causal discovery.

We assume that our detections (and the hierarchies used for such) are sufficient. In particular, the set of pre-specified actions is sufficient, and the computer is able to generally detect these elements in the scene when they occur.

We assume causal faithfulness: multiple causes do not exactly cancel. When we detect no correlation, we match this to the perception of no causal connection.

We assume each effect is a function of its immediate causes and an independent error. Each action, A_i depends on its own exogenous variable, u_{A_i}. Using ΔF_j to denote fluent change j, we notate in terms of structural equations:

$$A_i = g_{A_i}(u_{A_i}) \text{ for } i = 1, \ldots, n_A \tag{1}$$

$$\Delta F_j = g_{\Delta F_j}(\mathbf{A}_j, u_{\Delta F_j}) \text{ for } j = 1, \ldots, n_{\Delta F} \tag{2}$$

where $\mathbf{A}_j$ denotes specifically those actions that are in a causal relationship with ΔF_j. $u_{\Delta F_j}$ are exogenous.

3.3. Potential Effects: The Space of Fluent Changes

Given a fluent that can take n_F values, there are $n_{\Delta F} = n_F^2$ possible transitions from time t to $t+1$. With the door, for example, where the fluent could be "open" or "closed", there are four possible sequences: the door changes from "open" to "closed", changes from "closed" to "open", remains "open", or remains "closed". We notate the fluent change for a clip with ΔF.

Per the commonsense reasoning literature [Mueller 2006], a lack of change-inducing action (referred to here as "non"-action) causes the fluent to maintain its status, denoted $\Delta F = 0$; for example, a door that is closed will remain closed until some action changes that status. Figure 1 shows the door and the light maintaining their statuses for varied durations, punctuated by periods of change due to action.

The space of fluent changes possible over the objects in the video is pre-specified and denoted by

$$\Omega_{\Delta F} = \{\Delta F\}.$$

3.4. Potential Causes: The Space of Action Detections

Action parsing provides Ω_A, the space of actions. Ω_A contains actions at high levels of an action hierarchy. An action detection hierarchy (e.g., [Pei et al. 2011]) aggregates pixels into objects, relates these objects spatially and temporally to define atomic actions, groups those into sub-actions (such as pushing or pulling the door), and hierarchically combines even further (for example, unlocking and pulling the door). Figure 1 shows actions from different levels of the hierarchy.

In this paper, Ω_A is limited to top-level action or sub-actions from a pre-designed action hierarchy, following Heuristic 1.

4. PERCEPTUAL CAUSAL RELATIONS

In this section, we formalize our key building block for causal structure, the notion of a perceptual causal relation between an action and a fluent change.

4.1. Defining Perceptual Causal Relations

Combining the fluent changes with the actions, we define the space of potential causal relations.

Definition 4.1 (*Space of Causal Relations*). The space of causal relations is given by

$$\Omega_{CR} = \Omega_A \times \Omega_{\Delta F}. \tag{3}$$

The space, Ω_{CR}, provides the basic units for learning. Elements $\mathbf{cr} \in \Omega_{CR}$ specify an action and fluent change, and provide the framework for the 2×2 tables as shown in Table I.

Table I. Causal relation.

$\mathbf{cr}$:

	Action	¬Action
Effect	c_0	c_1
¬Effect	c_2	c_3

Labeling the individual cells of the table, $\mathbf{cr} = (c_0, c_1, c_2, c_3)$ where c_i functions as a binary indicator. When applied to a sufficiently short video clip (defined in the next section), the elements of Ω_{CR} identify whether or not the clip has the action and/or fluent change.

When these video clips show strong evidence for elements of Ω_{CR}, we award perceptual causal status and add the elements to our model.

4.2. Preparing the Data: Creating Clips from the Video

In order to determine the elements of Ω_{CR} which have the most evidence for being true causal relations, we evaluate the elements using video.

A long video sequence $\mathbf{V}$ is decomposed into shorter video clips, $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$. Following Heuristic 2 for limiting temporal lag, only actions occurring within a prespecified $\tau_{\max}$ of the fluent change are included in $\mathbf{v}_i$, to be considered as potential causes. The function $\tau(t_A, t_F)$ measures time between the action completion, t_A, and the fluent change, t_F. Some example functions for $\tau(t_A, t_F)$ include:

(1) Counting the number of frames between t_A and t_F. We consider $\tau_{\max}$ between 15 and 90 seconds.
(2) Counting the number of actions detected between t_A and t_F. We consider $\tau_{\max}$ ranging from 1 to 6 recent actions.
(3) Combinations of the first two. We consider $\tau_{\max}$ to be $\max$ or $\min$ over combinations of $15, 45$ seconds and $1, 2, 3$ actions. For example, taking the maximum of 15 seconds and 2 actions creates clips *at most* 15 seconds long or with *at most* 2 action detections. Taking the minimum of 15 seconds and 2 actions creates clips of *at least* 15 seconds or 2 action detections.

These are explored in experiments in Section 7.2.4. It is intuitive to expect a dependence between clip length definition and performance. If the clip is not long enough to include the causing action, then the ability to detect causes diminishes. However, if clip length is too long, then there will only be a few examples, not enough information to rise above the noise.

4.3. Evaluating Causal Relations

Aggregating the values from $\mathbf{cr} \in \Omega_{CR}$ across the clips, $\mathbf{v}_i$, we obtain relative frequencies for the particular action and fluent change:

Definition 4.2 (*Relative Frequencies of a Causal Relation*). Given a causal relation cr and video $\mathbf{V}$ that has been broken into clips $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, the relative frequencies of cr are given by

$$RF(\mathbf{cr}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{cr}(\mathbf{v}_i). \tag{4}$$

The relative frequencies from the video's action and fluent detections are denoted by $\mathbf{f} = (f_0, f_1, f_2, f_3)$.

Our causal model is built by greedily augmenting action and fluent distribution with causal relations, linking actions to fluent changes. At any iteration, there is the model that has been built so far (the "current model"), and the observed data from the video. The limiting relative frequencies under the current model are denoted by $\mathbf{h} = (h_0, h_1, h_2, h_3)$. Table II summarizes these statistics.

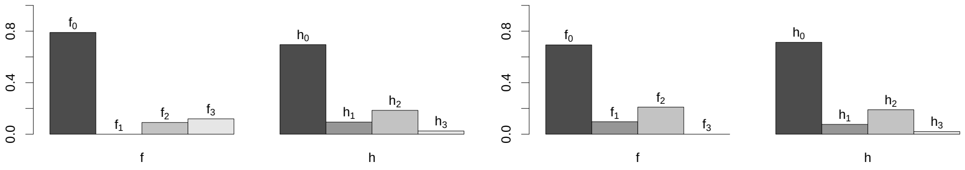
Table II. Relative Frequencies.

ΔF	A	Current Model	Observed Data
0	0	h_0	f_0
0	1	h_1	f_1
1	0	h_2	f_2
1	1	h_3	f_3

We construct our model by electing the most informative causal relations sequentially in terms of maximizing the information gain. Intuitively, this information gain is linked to the difference between $\mathbf{f}$ and $\mathbf{h}$.

For a causing action, $\mathbf{f}$ is shown in Figure 2(a), together with the relative frequencies of cr under a probability model assuming independence, $\mathbf{h}$. The greatest difference between these histograms occurs in the f_1/h_1 and f_3/h_3 components. The relative frequencies $\mathbf{f}$ and $\mathbf{h}$ for a non-causing action in (b) look equivalent, indicating independence between the fluent and action.

We select the relations that show the greatest difference between $\mathbf{f}$ and $\mathbf{h}$, as measured by the KL-divergence, thereby adding perceptual causal semantics to the model.



(a) Causing Action (b) Non-Causing Action

Fig. 2. Relative frequencies of **cr** for the observations is shown on the left of each pair, and for the model of independence on the right.

5. PURSUIT OF THE CAUSAL RELATIONS

In this section we develop the theoretical framework for the learning theory, formulas of which were provided in [Fire and Zhu 2013b]. From the space of all possible relations, Ω_{CR}, we now show how to sequentially select **cr** and build a joint probability model incorporating them.

The video clips, $\mathbf{v}_i$, are assumed to be drawn from an unknown distribution of perceptual causality, $f(\mathbf{v})$. We incrementally build a series of models approximating f

$$p_0(\mathbf{v}) \to p_1(\mathbf{v}) \to \ldots \to p(\mathbf{v}) \to p_+(\mathbf{v}) \to \ldots \to p_k(\mathbf{v}) \approx f(\mathbf{v}), \tag{5}$$

where each new model incorporates a new causal relation as illustrated in Figure 3. We use an information projection approach (see, e.g., [Csiszár and Shields 2004]).

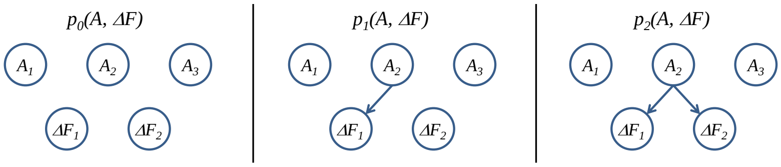


Fig. 3. The perceptual causal structure is incrementally constructed. Here, the action is flipping the light switch, which can turn the light on or off.

Shown in the first panel of Figure 3, learning initializes by independently considering action and fluent distributions, p_A and $p_{\triangle F}$, respectively:

$$p_0(\mathbf{v}) = p_A(\mathbf{v}) p_{\triangle F}(\mathbf{v}). \tag{6}$$

In this paper, we initialize $p_A(\mathbf{v})$ with the proportion of clips, $\mathbf{v}$, that contain action A; similarly for $p_{\triangle F}$.

In a single iteration, we fix the previous model, p, and augment to a new model, p_+. Under the information projection framework, learning proceeds in two steps. In the first step, we select the causal relation to add to the model by maximizing the KL-divergence between p_+ and p, also known as the information gain. In step two, we fit the selected causal relation to the data by minimizing the KL-divergence between p_+ and p.

Any model over the video clips that considers fluent changes independently from causing actions, such as p_0, will fail to match f on true causal relations. However, given a selected relation, the latter step requires that the new model match the observed data on the newly selected causal relation

$$E_{p_+}[\mathbf{cr}_+] = E_f[\mathbf{cr}_+] \approx \mathbf{f}. \tag{7}$$

The probability distribution with minimum KL-divergence, $\mathrm{KL}(p_+||p)$, subject to that constraint is

$$p_+(\mathbf{v}) = \frac{1}{z_+}p(\mathbf{v})\exp\left(-\langle\lambda_+, \mathbf{cr}_+\rangle(\mathbf{v})\right) \tag{8}$$

where $\lambda_+ = (\lambda_0, \lambda_1, \lambda_2, \lambda_3)$ is a scalar vector corresponding to the components of $\mathbf{cr}_+(\mathbf{v}) = (c_0(\mathbf{v}), c_1(\mathbf{v}), c_2(\mathbf{v}), c_3(\mathbf{v}))$ shown in Table I and described in Section 4.1 and z_+ is a normalizing constant. When p_0 is uniform, this procedure yields the maximum entropy distribution.

5.1. Fitting the Causal Relation

Unlike in other information projection applications to vision (e.g., [Della Pietra et al. 1997] [Zhu et al. 1997]), λ_+ can be computed analytically thanks to the binary nature of the causal relation:

PROPOSITION 5.1. *To add the causal relation* $\mathbf{cr}_+$ *to the model in Equation 8, the parameters are given by:*

$$\lambda_i = \log\left[\frac{h_i}{h_0}\cdot\frac{f_0}{f_i}\right] \tag{9}$$

for $i = 0, \ldots, 3$*, where* h_i *and* f_i *are as found in Table II.*

PROOF OF PROP. 5.1. Consider adding a single causal relation to the probability distribution, $p(\mathbf{v}) = \frac{1}{Z}exp(-\mathcal{E}(\mathbf{v}))$. This gives a new probability distribution

$$p_+(\mathbf{v}) = \frac{1}{z_+}p(\mathbf{v})\exp\left(-\left\langle\lambda_+, \mathbf{cr}_+(\mathbf{v})\right\rangle\right). \tag{10}$$

Since $\sum_{i=0}^{3} c_i = 1$, there is 1 degree of freedom in λ_+; without loss of generality, set $\lambda_0 = 0$.

From the observed data, the expected value under the true distribution, f, is best estimated by the quantity from the data,

$$E_f(c_i(\mathbf{v})) = f_i. \tag{11}$$

Further, $E_p(c_i(\mathbf{v})) = h_i$.

$$\begin{aligned} E_{p_+(\mathbf{v})}(c_i(\mathbf{v})) &= \int p_+(\mathbf{v})c_i(\mathbf{v})d\mathbf{v} \quad (12)\\ &= \int \frac{1}{z_+}p(\mathbf{v})\exp(-\langle\lambda_+, \mathbf{cr}_+(\mathbf{v})\rangle)c_i(\mathbf{v})d\mathbf{v} \quad (13)\\ &= E_p\left(\frac{1}{z_+}\exp(-\langle\lambda_+, \mathbf{cr}_+(\mathbf{v})\rangle)c_i(\mathbf{v})\right) \quad (14)\\ &= \frac{1}{z_+}h_i\exp(-\lambda_i) \quad (15) \end{aligned}$$

The last equation holds because only one of the $c_i(\mathbf{v})$ will be nonzero at a time. Equating the matched statistics,

$$f_i = \frac{1}{z_+}h_i\exp(-\lambda_i). \tag{16}$$

Since $\lambda_0 = 0$, $f_0 = \frac{h_0}{z_+}$, or

$$z_+ = \frac{h_0}{f_0}. \tag{17}$$

Hence,

$$\lambda_i = \log\left[\frac{h_i}{h_0} \cdot \frac{f_0}{f_i}\right]. \tag{18}$$

□

Intuitively, the h_i/h_0 component "undoes" the independent consideration under the current model, and the f_0/f_i component inserts the new information joining the action and fluent change.

In experiments, $p_0(\mathbf{v})$ is defined over a finite set, and $\mathbf{h}$ is computable.

5.2. Pursuing Causal Relations by Information Projection

While Proposition 5.1 provides a formula to add a causal relation to a model, the best causal relation, $\mathbf{cr}_+$, is selected at each step through a greedy pursuit which leads to the maximum reduction of the KL divergence [Della Pietra et al. 1997], [Zhu et al. 1997]:

$$\mathbf{cr}_+ = \operatorname*{argmax}_{\mathbf{cr}} \left(\mathrm{KL}(f||p) - \mathrm{KL}(f||p_+)\right). \tag{19}$$

Equivalently, $\mathbf{cr}_+$ is added to maximize the information gain:

$$\mathbf{cr}_+ = \operatorname*{argmax}_{\mathbf{cr}} IG_+ := \operatorname*{argmax}_{\mathbf{cr}} \mathrm{KL}(p_+||p) \geq 0, \tag{20}$$

moving the model closer to the true distribution f with each new causal relation.

An analytic formula provides the best causal relation:

PROPOSITION 5.2. *The next best relation,* $\mathbf{cr}_+$*, to add to the model is given by*

$$\mathbf{cr}_+ = \operatorname*{argmax}_{\mathbf{cr}} \mathrm{KL}(p_+||p) = \operatorname*{argmax}_{\mathbf{cr}} \mathrm{KL}(\mathbf{f}||\mathbf{h}) \tag{21}$$

where $\mathbf{f}$ *and* $\mathbf{h}$ *are as found in Section 4.3.*

PROOF OF PROP. 5.2.

$$\begin{aligned}
\mathrm{KL}(p_+||p) &= \int p_+(\mathbf{v}) \log \frac{p_+(\mathbf{v})}{p(\mathbf{v})} d\mathbf{v} && (22)\\
&= \int p_+(\mathbf{v}) \log\left(\frac{1}{z_+}\exp(-\langle\lambda_+, \mathbf{cr}_+(\mathbf{v})\rangle)\right) d\mathbf{v} && (23)\\
&= \int p_+(\mathbf{v}) \log \frac{1}{z_+} d\mathbf{v} - \int p_+(\mathbf{v})(\langle\lambda_+, \mathbf{cr}_+(\mathbf{v})\rangle) d\mathbf{v} && (24)\\
&= \log \frac{1}{z_+} - E_{p_+}(\langle\lambda_+, \mathbf{cr}_+(\mathbf{v})\rangle) && (25)\\
&= \log \frac{1}{z_+} - E_f(\langle\lambda_+, \mathbf{cr}_+(\mathbf{v})\rangle) && (26)\\
&= \log \frac{1}{z_+} - \langle\lambda_+, \mathbf{f}\rangle. && (27)
\end{aligned}$$

Applying the formula for λ_i,

$$\begin{aligned}
\lambda_i f_i &= f_i \log \left[\frac{h_i}{h_0} \cdot \frac{f_0}{f_i} \right] && (28) \\
&= f_i \log \frac{f_0}{h_0} + f_i \log \frac{h_i}{f_i}. && (29)
\end{aligned}$$

Continuing from Equation 27 and substituting Equations 17 and 29,

$$\begin{aligned}
\mathrm{KL}(p_+ || p) &= \log \frac{f_0}{h_0} - \sum_{i=0}^{3} \left(f_i \log \frac{f_0}{h_0} + f_i \log \frac{h_i}{f_i} \right) && (30) \\
&= (1 - f_1 - f_2 - f_3) \log \frac{f_0}{h_0} + \sum_{i=1}^{3} f_i \log \frac{f_i}{h_i} && (31) \\
&= f_0 \log \frac{f_0}{h_0} + \sum_{i=1}^{3} f_i \log \frac{f_i}{h_i} && (32) \\
&= \mathrm{KL}(\mathbf{f} || \mathbf{h}). && (33)
\end{aligned}$$

□

Therefore, in order to determine which causal relation is best to add to the model, we calculate the KL-divergence between the current model and the data for each potential causal relation, selecting the one that maximizes the information gain.

Once the relation is selected, perceptual causal arrows can be assigned between A and ΔF according to Heuristic 1 as shown in Figure 3.

Algorithm 1 summarizes Propositions 5.1 and 5.2.

ALGORITHM 1: Learning the causal relations.

```
Input  : Action and fluent change detections from the video, $\tau(t_A, t_F)$ and $\tau_{\max}$
Output: Probability distribution over a learned structure of perceptual causality
Create video clips according to $\tau$ and $\tau_{\max}$;
Tally observations;
Initialize model estimates (e.g., with proportions of action/fluent change occurrence);
repeat
    foreach candidate causal relation do
        Compute its information gain by Proposition 5.2;
    end
    Select cr that maximizes information gain;
    Calculate $\lambda_+$ by Proposition 5.1;
    Update model estimates using $\lambda_+$;
until information gain is smaller than a threshold;
```

5.3. Precise Selection of cr When Actions are Hierarchical

In recent computer vision literature, human actions are organized into hierarchical representations, such as stochastic event grammar [Ivanov and Bobick 2000] or the Temporal And-Or Graph [Pei et al. 2011]. In such representations, actions can be decomposed into sub-actions (where all parts compose the action) and alternative actions. The Temporal And-Or Graph represents these as And-nodes and Or-nodes, respectively.

As instances of a parent and its children often compete, our learning method must have the precision to select the correct node as the cause of the fluent change. Fortunately, as the information gain for each action node in the action hierarchy is tested, these parent/child interactions are automatically taken into account.

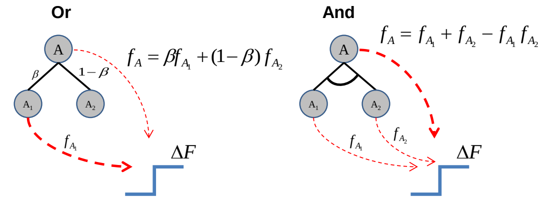

Or
A
β
$1-\beta$
A₁
A₂
$f_A = \beta f_{A_1} + (1-\beta) f_{A_2}$
f_{A_1}
ΔF
And
A
$f_A = f_{A_1} + f_{A_2} - f_{A_1} f_{A_2}$
A₁
A₂
f_{A_1}
f_{A_2}
ΔF

Fig. 4. Graphical demonstration for when our algorithm encounters an Or-node or an And-node in the action hierarchy. When encountering an Or-node, where action A is identified through one of the child actions A_1 or A_2 with prior probability of A_1 of β, the pursuit process prefers the child node showing the strongest causal relation. For an And-node, where the action A is identified as a composition of A_1 and A_2, the parent is preferred.

We first consider a parent action, A, that is a choice between two children, A_1 or A_2, as shown with the Or-node on the left of Figure 4. Intuitively, if A_1 is a cause, but not A_2, then A_1 will exhibit the strongest relationship with the fluent change. A will have the second highest, as some of the time it is activated when A_1 occurs and some of the time it is activated when A_2 occurs.

For a cause, the information gain is dominated by the $f_3 \log f_3/h_3$ contribution. Let f_A, f_{A_1}, and f_{A_2} be f_3 from Table II for A, A_1, and A_2, respectively. Further, let β be the Or-probability of selecting A_1. In this case,

$$f_A = \beta \cdot f_{A_1} + (1-\beta) \cdot f_{A_2}, \tag{34}$$

and therefore,

$$\min(f_{A_1}, f_{A_2}) \leq f_A \leq \max(f_{A_1}, f_{A_2}). \tag{35}$$

Further, let h_A, h_{A_1}, and h_{A_2} be defined similarly. Since A happens if A_1 or A_2 happen, $h_A > h_{A_1}$.

Finally, if $h_3 < f_3$ as is the case on a distribution considering A and ΔF independently, then

$$h_{A_1} < h_A < f_A \leq f_{A_1}, \tag{36}$$

and the contribution on the information gain for A_1 will be larger than for A. In the case of an Or-node, the causing child node will be selected over the parent under pursuit by information gain.

Next, let A be a parent that groups its children A_1 and A_2 as in the right side of Figure 4. In this case, A happens if both children A_1 and A_2 happen and so $h_A < h_{A_1}$ and

$$f_A = f_{A_1} + f_{A_2} - f_{A_1} f_{A_2}. \tag{37}$$

It follows that

$$f_A \geq f_{A_1}, f_{A_2} \tag{38}$$

and

$$h_A < h_{A_1} < f_{A_1} \leq f_A. \tag{39}$$

Therefore, for an And-node where both children must happen in order for the parent node to happen, our method selects the parent node.

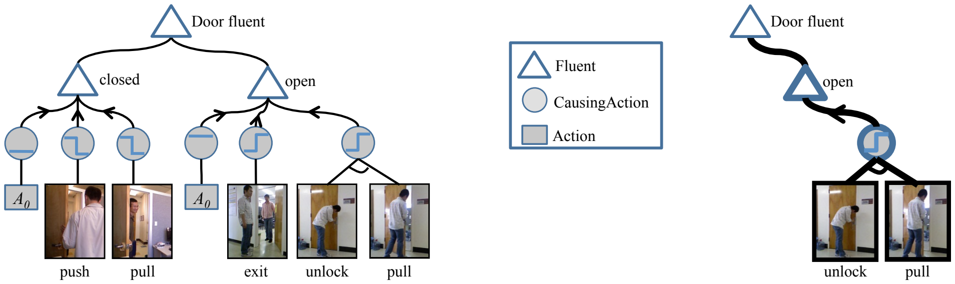


Fig. 5. The Causal And-Or Graph (left) and a parse graph (right). Each causing action node shows an action from a high level of the hierarchy. Arrows point from these actions (causes) to the fluent (effect). Children of And-nodes are grouped by arcs. A_0 represents non-action, causing a fluent to maintain status.

6. THE CAUSAL AND-OR GRAPH

The And-Or structure that is used to represent actions was first introduced to computer vision for representing compositional concepts defined by grammar [Zhu and Mumford 2006], and it can be used to collect the perceived causal relations. An example of the Causal And-Or Graph for a fluent value [Fire and Zhu 2013b] is shown in Figure 5. Nodes are hierarchically defined: And-nodes activate if all their children activate, while Or-nodes activate if any of their children activate.

The Causal And-Or Graph provides a detangled view of the causal structure, separating the causes. Or-nodes represent fluent values, whose children are the alternate causes for that fluent value. For example, ΔF as an Or-node could be caused by any of the alternative causes A_1, A_2, and A_3:

$$\Delta F \leftarrow A_0 \vee A_1 \vee A_2. \tag{40}$$

Arrows point from causes to effects. These Or-nodes represent a choice in the causing condition. Here, actions cause fluent values to change. Similarly, non-actions (shown in Figure 5 with A_0) maintain a fluent's value.

Action recognition, while beyond the scope of this paper, works by detecting spatio-temporal relationships in the video (e.g., detecting computer use through relative positions of skeleton joints and proximity to the computer [Wei et al. 2013]). These spatio-temporal relationships are really compositions of fluents (as ambient conditions or as the visual decomposition of actions). In the Causal And-Or Graph, these compositions are represented with And-nodes, e.g.,

$$A_2 := f_1 \wedge f_2 \tag{41}$$

where := represents definition.

The Causal And-Or Graph provides a hierarchical, computationally efficient decomposition that is useful in computer vision for detections. A selection on the Or-nodes provides parse graphs (pg) from the grammar and represents simpler causal explanations; an example of which is shown in the left side of Figure 5. This parse graph provides the causal explanation for the video clip: the door is open because an agent unlocked and pulled. These untangled networks allow faster inference.

Further, the Or-nodes encode *prior* information on the different causes. (This is different from Bayesian structural equation modeling, which places a prior over the parameters [Scheines et al. 1999].) Humans have an intuitive understanding of causation that they use to answer questions amid missing or hidden information. Without seeing what happened or knowing what the circumstances are in the room, they can answer: Why is the door closed? (Because no one opened it.) Why did the light turn on?

(Because someone toggled the switch.) A prior on causality is important for computer vision as it enables guesses on the particular causal relationship (both the cause and effect together) in play when only partial information is available and thus can fill in detections.

Note that the learned Causal And-Or Graph depends on both the pre-specified fluents of interest and the action recognition hierarchy used. For example, here we learned the joint actions of unlock and push open the door. This could more accurately be represented by changing the lock's fluent, coupled with the pushing action. Regardless, the learning method still produces a graph structure that is useful.

6.1. The Probability Model on the Causal And-Or Graph

The Causal And-Or Graph is a graphical representation of the joint probability distribution learned in Section 5, conditioned on the fluent value. This natural probability distribution over the Causal And-Or Graph provides the prior on causality.

More concretely, probability is defined over the parse graphs, pg, in the Causal And-Or Graph, and is formed by conditioning on the fluent value in the jointly pursued model:

$$p_{\mathrm{C}}(pg) = p(pg|F) \propto \exp\left(-\mathcal{E}_{\mathrm{C}}(pg)\right) \tag{42}$$

where

$$\mathcal{E}_{\mathrm{C}}(pg) \ = \mathcal{E}_0(pg) + \sum_{a \in CR(pg)} \lambda_a(w(a)). \tag{43}$$

$\mathcal{E}_0(pg)$ is the energy from the model p_0 in Equation 6, limited to the actions and fluents relevant to the included causal relations. $CR(pg)$ is the set of all non-empty, causal relations included in the parse graph (Or-nodes). $w(a)$ is the choice of causing action a (the selection of the child from the Or-node). λ_a comes from Equation 5.1 and represents the switch probability on the Or-nodes for cr_a, providing a measure for how frequently an action causes the fluent status.

This prior on causality allows common knowledge to overcome ambiguous or missing spatio-temporal detections. When this prior distribution over the parse graphs is combined with a likelihood model, MAP inference provides instances of perceptual causality in video.

This probability on the Causal And-Or Graph can be thought of as a scoring mechanism for detection purposes. In particular, detections of fluents and actions contribute to the score, and the prior on causality contributes a favorable amount to the score if the actions and fluents detected are linked.

7. EXPERIMENTS

In this section, we apply the learning theory developed in Section 5. Our model for perceptual causality is evaluated against human perception.

7.1. Simulated Vending Machine

To test the learning process amid incorrect action detections, we simulated a vending machine with the joint Spatio-Temporal-Causal And-Or Graph shown in Figure 6. An agent can use the vending machine or perform a confusing action. Using the vending machine correctly causes the machine to vend various confections. Some example sequences synthesized from the graph are:

1	Arrive, Push D, Push 1, Leave.
2	Arrive, Pay, Push A, Push 1, Get Snack, Leave. Machine Vends Chocolate.

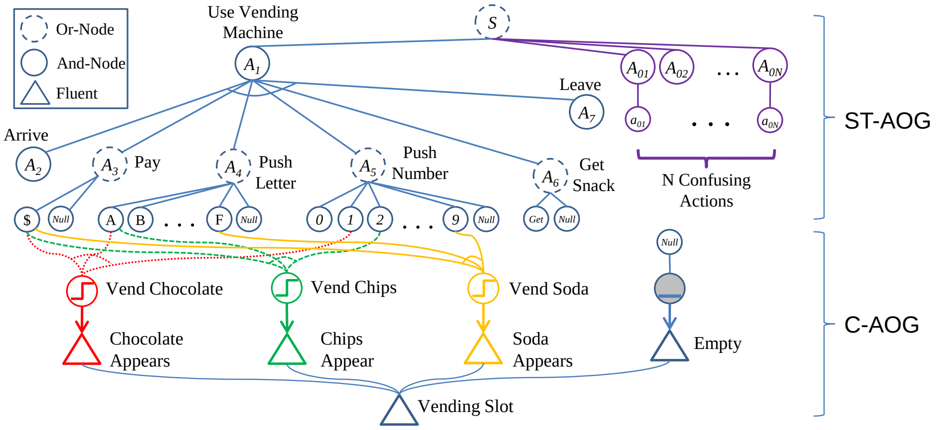
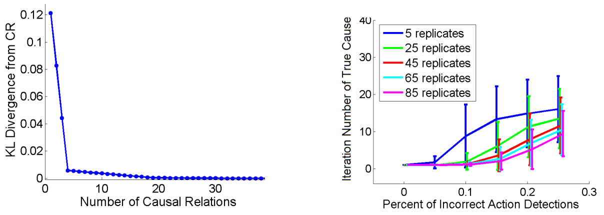


Fig. 6. Spatio-Temporal-Causal And-Or Graph for simulated vending machine. To use the vending machine, a code must be entered using an alphanumeric pad. With payment, the correct combination will cause the machine to vend one of three snacks: chips, chocolate, or soda. With an incorrect code or no payment, the vending slot remains empty.

Individual nodes, including 10 confusing actions and combinations thereof, are considered as potential causes for the machine to vend the various confections. The KL-divergence between the true data and the learned model that is attributable to causal relations is shown in Figure 7(a). After learning the true causal relations, the model learns noise, but these causal relations contribute minimally to the reduction in KL-divergence and are not generalizable.



(a) KL-divergence as causal relations are added to the model.

(b) Iteration in which true cause is selected when varying the number of misdetections.

Fig. 7. Simulation results.

We randomly change a fraction ($p = 0, 0.05, 0.1, 0.15, 0.2, 0.25$) of simulated actions and fluents to provide noise that would occur with detection algorithms.

Possibilities from the And-Or Graph are sampled $N = 5, 25, 45, 65, 85$ times, creating replicates. The number of iterations to detect the true cause is calculated. Results are shown in Figure 7(b) where error bars are estimated using 500 different samples of each replicate. Under replication of the experiment design, our methods are able to overcome faulty action detection, ranking the true cause appropriately.

7.2. Learning Multiple Causal Relations Amid Confusing Actions

7.2.1. Video Data. A video was recorded with a Kinect sensor in an office scene. Actions in the scene are listed in Table III. Fluents include door open/closed, light on/off, and computer monitor on/off. The Causal And-Or Graph of Figure 8 shows some screenshots of the video. The video contains 8 to 20 (sometimes simultaneous) instances of each action category. There are a total of 66 possible action-fluent relations, with 10 true causal relationships among them.

Table III. Legend of actions for office scene.

A_i	Description
A_0	Non-action, no explaining action
A_1	Open the door from the inside
A_2	Close the door from the inside
A_3	Open the door from the outside
A_4	Close the door from the outside
A_5	Touch the power button on the monitor
A_6	Touch the mouse
A_7	Touch the keyboard
A_8	Touch the light switch
A_9	Confusing action: pick something up
A_{10}	Confusing action: have a conversation
A_{11}	Confusing action: walk by

In this office scene experiment, we start with perfect action and fluent detections to demonstrate learning. We compare these results to those obtained with noisy detections.

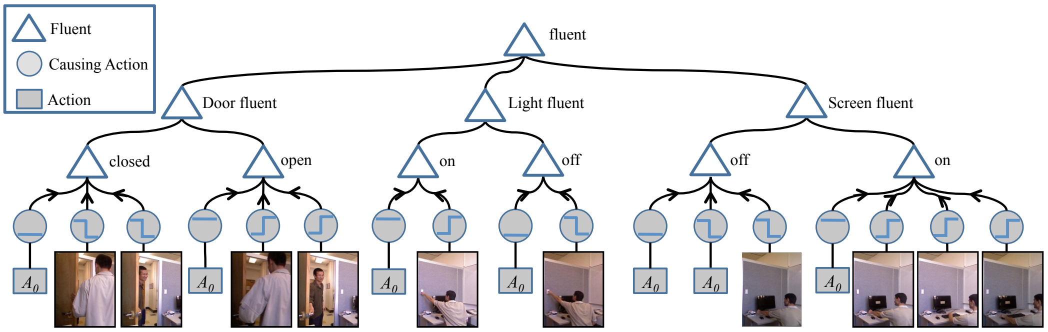


Fig. 8. A Causal And-Or Graph for door status, light status, and screen status. Action A_0 represents non-action (a lack of state-changing agent action). Non-action is also used to explain the change of the monitor status to off when the screensaver activates. Arrows point from causes to effects, and undirected lines show deterministic definition.

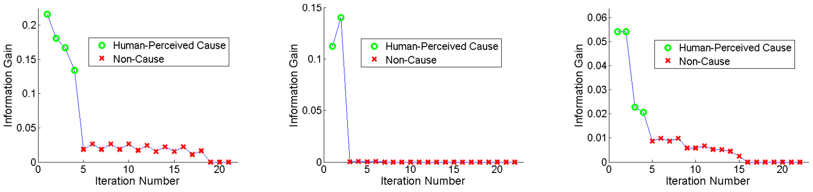
Table IV shows information gains during the pursuit process for the door fluent. In the first 4 iterations, all four correct causal relations are selected. Once the relation has been fit, the model does not gain information for that relation.

Figure 9 shows plots of information gains for causal relations in the order pursued, separated by fluent. Causes are added to the model before non-causes. Clear cutoffs of information gains for the door and light fluents separate causes from non-causes.

The correct cutoff is less clear for the computer monitor, in part due to only acquiring partial causal information. The monitor's display status has preconditions of power and computer status which were not detectable.

Table IV. Information gains for the top 20 causal relations involving the door fluent (columns) over 15 iterations (rows). The highest information gain in each iteration is shown bolded. True causes are shown with a gray background.

	$C\to O$	$O\to C$	$O\to C$	$C\to O$	$O\to C$	$C\to O$	$O\to C$	$C\to O$	$O\to C$	$C\to O$	$O\to C$	$C\to O$	$O\to C$	$C\to O$	$O\to C$	$C\to O$	$O\to C$
	A_3	A_4	A_2	A_1	A_6	A_6	A_7	A_7	A_8	A_8	A_{10}	A_{10}	A_5	A_5	A_9	A_9	A_{11}
$k=1$	**0.2161**	0.1812	0.1668	0.1344	0.0185	0.0185	0.0185	0.0185	0.0185	0.0185	0.0170	0.0170	0.0155	0.0155	0.0155	0.0155	0.0111
$k=2$	0.0000	**0.1812**	0.1668	0.1344	0.0185	0.0185	0.0185	0.0185	0.0185	0.0185	0.0170	0.0170	0.0155	0.0155	0.0155	0.0155	0.0111
$k=3$	0.0000	0.0000	**0.1668**	0.1344	0.0185	0.0185	0.0185	0.0185	0.0185	0.0185	0.0170	0.0170	0.0155	0.0155	0.0155	0.0155	0.0111
$k=4$	0.0000	0.0000	0.0000	**0.1344**	0.0185	0.0185	0.0185	0.0185	0.0185	0.0185	0.0170	0.0170	0.0155	0.0155	0.0155	0.0155	0.0111
$k=5$	0.0000	0.0000	0.0000	0.0000	**0.0185**	**0.0185**	**0.0185**	**0.0185**	0.0185	0.0185	0.0170	0.0170	0.0155	0.0155	0.0155	0.0155	0.0111
$k=6$	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0264**	0.0185	0.0185	0.0185	0.0185	0.0170	0.0170	0.0155	0.0155	0.0155	0.0155	0.0111
$k=7$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0185**	**0.0185**	0.0185	0.0185	0.0170	0.0170	0.0155	0.0155	0.0155	0.0155	0.0111
$k=8$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0264**	0.0185	0.0185	0.0170	0.0170	0.0155	0.0155	0.0155	0.0155	0.0111
$k=9$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0185**	**0.0185**	0.0170	0.0170	0.0155	0.0155	0.0155	0.0155	0.0111
$k=10$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0264**	0.0170	0.0170	0.0155	0.0155	0.0155	0.0155	0.0111
$k=11$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0170**	**0.0170**	0.0155	0.0155	0.0155	0.0155	0.0111
$k=12$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0244**	0.0155	0.0155	0.0155	0.0155	0.0111
$k=13$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0155**	**0.0155**	**0.0155**	**0.0155**	0.0111
$k=14$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0224**	0.0155	0.0155	0.0111
$k=15$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0155**	**0.0155**	0.0111



(a) Door (b) Light (c) Monitor

Fig. 9. Information gains for causal relations in the order pursued, separated by fluent. Green circles label causes.

7.2.2. Comparisons: Hellinger's χ^2 and TE. As learning causal structure is new to vision research, there are no benchmarks for comparison. Instead, we compare our learning technique to ranks of causal effects and to measurements of independence.

Potential causes can be ranked based on their causal effect. One such measure is the treatment effect, TE, of treatment A over $\neg A$:

$$TE = E(\Delta F | do(A)) - E(\Delta F | do(\neg A)). \tag{44}$$

The larger $|TE|$ is, the stronger the causal effect.

As a further comparison tool, one standard measurement of independence is the χ^2 statistic. Due to low expected cell frequencies, the standard χ^2 measure is insufficient. Instead, we compare our results to Hellinger's χ^2, a more robust measure.

On this experiment, our results are validated with similarly ranked values of TE and χ^2.

7.2.3. Noisy Data. Randomly changing different percentages of action detections leads to the curves shown in Figure 10. As more noise enters the system, the information gained by considering causal relations decreases. While learning works amid noisy scenes (many actions happening simultaneously), clean detections are important.

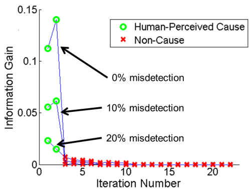
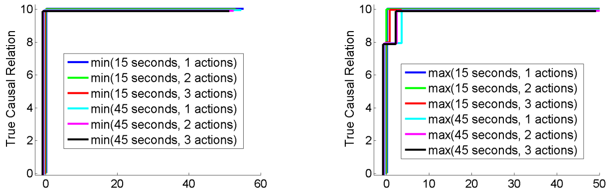


Fig. 10. Information gains for causal relations in the order pursued, separated by fluent.

7.2.4. Varying Video Clip Length: The Effect of $\tau()$ and $\tau_{\max}$. This experiment explores the choice of $\tau()$ and $\tau_{\max}$ as described in Section 4.2, with simultaneous pursuit of door, light, and monitor causal relations.

A video may show periods of clutter with many actions happening at once, whereas other times show no actions at all. We take $\tau()$ as a minimum (or maximum) over two methods for measuring time (counting seconds and counting actions). This ensures an example has a short duration if nothing is happening while simultaneously limiting the number of actions considered. Figure 11 highlights that the minimum outperforms the maximum.

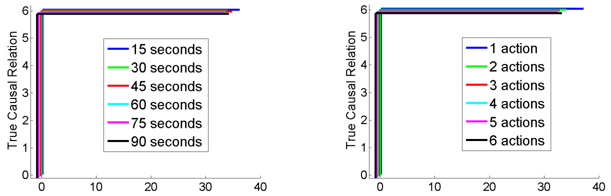
The longer the time span used to build an example of the desired fluent, the more confusing actions enter as potentially relevant. Keeping the number of considered actions small makes the examples cleaner, decreasing noise obscuring the causal links. Optimally, the timespan used will be short, but special attention is required when considering events subject to a time delay.



(a) Minimum (b) Maximum

Fig. 11. ROC curves for the joint pursuit of door, light, and monitor causal relations. Ten total causal relations.

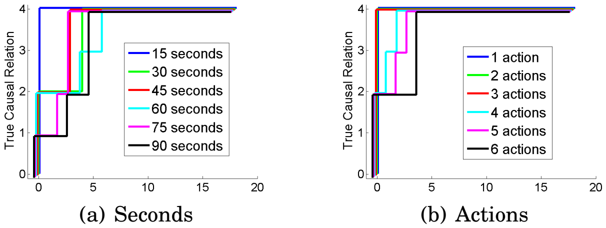
Focusing on the clear causal relations of the door and the light, Figure 12 shows their causal relations are 100% detectable when constructing examples using a fixed number of seconds (a) or a fixed number of actions (b).



(a) Seconds (b) Actions

Fig. 12. ROC curves for joint pursuit of the door and light causal relations. Six total causal relations.

Because the monitor's power both confounds the causal effects of the keyboard and mouse, and is a cause itself, detecting all causal relations for the monitor is difficult, as shown in Figure 13. The learning process sees some examples where these actions lead to the fluent change and some where they do not, but there are no cues to differentiate between those cases. Lower TE and χ^2 reflect the confusion in detecting the causal status of the power button.

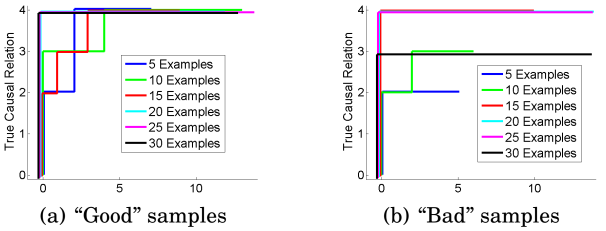


(a) Seconds (b) Actions

Fig. 13. ROC curves for the monitor. Four total causal relations.

7.2.5. Number of Examples Needed to Perceive Causal Structure. Reducing the number of examples used to learn causal relations has a detrimental effect on detection. Taking N random samples from the 97 examples, Figure 14(a) shows that as the number of examples used in training decreases, the ability to detect causal relations for the door fluent also decreases.

Figure 14(b) emphasizes the importance of quality, not quantity, in examples. While causes were recoverable in (a) with 5 examples, causes will never be recovered under the sample of 30 examples in (b). To identify a cause, there must be positive and negative examples.



(a) “Good” samples (b) “Bad” samples

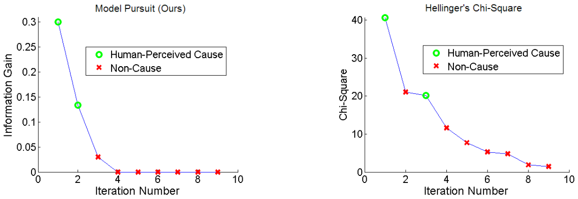
Fig. 14. ROC curve using N randomly selected examples to determine causal relations for the door fluent.

7.3. Experiments on Detections from Video

To validate performance against real data, experiments in this section use fluent change and action detections from video captured with the Kinect camera. Fluent changes were detected using the GentleBoost algorithm [Friedman et al. 2000] on a 3-level spatial pyramid [Lazebnik et al. 2006]. Actions were detected using relative joint positions of the skeletons output from the Kinect [Wei et al. 2013], coupled with the Temporal And-Or Graph [Pei et al. 2011].

7.3.1. Hierarchical Action Selection and Hellinger's χ^2. Where compound actions (in the doorway scene, unlocking with a key or entering a code, followed by pushing/pulling the door; opening from the interior of the room) are required for the effect, the causing actions may come from any level of the action hierarchy. Figure 5 shows the learned Causal And-Or Graph for the doorway scenes.

Our method maintains dependencies for actions that occur together; actions related to each other are suppressed once the cause is selected. Figure 15 shows that Hellinger's χ^2 fails to identify the correct causes, unable to suppress the dependence between hierarchically-related actions once a parent (or child) action is selected.

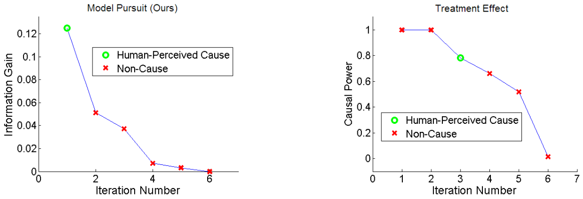


(a) Our Method (b) Hellinger's χ^2

Fig. 15. Hierarchical Example: The Locked Door, pursuit order of causes.

7.3.2. Delayed Effects and TE. This experiment uses detections from video of an elevator waiting area. For an elevator, the only detectable causing action to open the door is pushing the button that calls the elevator.

In this example, our method outperforms the treatment effect, TE, (Eq. 44) as shown in Figure 16.



(a) Our Method (b) TE

Fig. 16. Confounded Example: The Elevator, pursuit order of causes.

In this scenario, for all 4 times that someone walked away, the elevator doors opened (because they had first pushed the call button). As a measure, the treatment effect favors relationships when an action co-occurs with a fluent change 100% of the time—regardless of how infrequently the relationship is observed. Of the 19 total instances of opening doors, only 16 occur with the pushing button action under clip construction. Our method, however, incorporates the frequency with which the relationship occurs by examining the full contingency table.

7.4. Reasoning in Surprising Circumstances

Answers for "why" queries are obtained using MAP estimation. Observing a person pushing on the door while another agent walks by and yet another picks an object up, the learned probability model returns the correct reason for why the door is open.

If the door opens spontaneously (i.e., in a manner not seen by the system during learning), the probability model on the Causal And-Or Graph resolves the discrepancy by juggling the prior against which detection is more likely to be incorrect: the fluent change or the lack of action.

During the learning process for the monitor, however, the system saw several unexplained examples (i.e., when the computer put the monitor to sleep after sufficient time). In this case, the system learned to explain the status through the unexplained change, awarding 12% maximum posterior probability to the spontaneous change when no action is detected for turning the screen off.

8. SUMMARY AND DISCUSSION

Causal knowledge is required to fully explain the content of image and video data from an agentive point-of-view. In this paper, we have provided a learning framework for the perceptual causal structure between actions and fluents in video.

Causal relations were incrementally determined using the information projection principle, and we provided analytic formulas for selecting and adding the best causal relation to the current probability model. The information projection framework presented allows perceptual causal knowledge to be learned alongside actions and objects under other information projection frameworks, where information gains can be compared, and, for example, an important causal link could be added to a model before a less significant object or action.

The learned Causal And-Or Graph aligns with forms used in vision for detecting actions, objects, and fluents, and flattens a causal network into choices. The Or-nodes place a prior on causality, to deal with the ambiguities of detections in vision. Detection probabilities are evaluated alongside the prior probability for the causal explanation.

Our method was validated against human perception, and produced a better causal structure than TE and Hellinger's χ^2-statistic. It has the precision to select the correct action from a hierarchy, where a parent action may explain a fluent change better than any of its children actions separately or vice versa.

General causal networks were too vague for our purposes. Cognitive science informed what variables (and at what level in the hierarchy) to consider as causes and effects, how to partition a long video into "examples", and when to causally relate actions and fluents.

Any assumptions exclude possible cases. We were unable to consider confounders, such as monitor power status, because our pre-specified action hierarchy excluded that interaction.

Even amid true causal sufficiency, we still are unsure which causal questions are important. Is it that the person is typing on the keyboard, or that the person is typing their password?

The detection error we expect in a vision system complicates studying: classifiers are not perfect; misdetection, occlusion, and bad data are common problems. How can we tell the difference between a true confounding cause versus noisy data? Some of these misdetection problems might be inherent to the system, disallowing independent exogenous variables.

A minor point on detection error: the detected cause may not be considered complete before the detection of the effect is begun. One way around this problem is to compare the start time of the cause against the end time of the effect, but this could have stronger implications on the temporal lag considered. As we showed in experiments, large lag obfuscates the causal relationships with so few examples.

With a static surveillance camera, you might not have both positive and negative examples of each action considered. In future work, we will also explore dynamic experimental design to determine what on-camera interventions would best confirm (or refute) the model's belief about causal relationships.

Even with its problems, perceptual causality allowed us to construct a working model of causality from video. And maybe getting things wrong is ok: if a human repeatedly perceived something, they would still form a model based on that. The model may not be correct, but it yields useful results. One area for future research is to adapt the model as new "surprising" information comes in, as a human would.

ACKNOWLEDGMENTS

We would like to thank Judea Pearl and Elias Bareinboim for discussions on causality and Ying Nian Wu for discussions on social statistics. We would also like to thank Mingtao Pei for assisting with data collection and providing event parsing results used in this project.

REFERENCES

M. Albanese, R. Chellappa, N. Cuntoor, V. Moscato, A. Picariello, VS Subrahmanian, and O. Udrea. 2010. Pads: A probabilistic activity detection framework for video data. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 12 (2010), 2246–2261.

M. Brand. 1997. The "Inverse Hollywood Problem": From video to scripts and storyboards via causal analysis. In *Proceedings of the National Conference on Artifial Intelligence*. 132–137.

S. Carey. 2009. *The origin of concepts*. Oxford University Press.

I. Csiszár and P. C. Shields. 2004. Information theory and statistics: A tutorial. *Communications and Information Theory* 1, 4 (2004), 417–528.

S. Della Pietra, V. Della Pietra, and J. Lafferty. 1997. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 4 (1997), 380–393.

A. Fire and S.-C. Zhu. 2013a. Learning Perceptual Causality from Video. In *AAAI Workshop: Learning Rich Representations from Low-Level Sensors*.

A. Fire and S.-C. Zhu. 2013b. Using causal induction in humans to learn and infer causality from video. Proceedings of the 35th Annual Conference of the Cognitive Science Society, 2297–2302.

J. Friedman, T. Hastie, and R. Tibshirani. 2000. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The annals of statistics* 28, 2 (2000), 337–407.

T.L. Griffiths and J.B. Tenenbaum. 2005. Structure and strength in causal induction. *Cognitive Psychology* 51, 4 (2005), 334–384.

York Hagmayer and Michael R Waldmann. 2002. How temporal assumptions influence causal judgments. *Memory & Cognition* 30, 7 (2002), 1128–1137.

D. Heckerman. 1995. A Bayesian approach to learning causal networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 285–295.

Y.A. Ivanov and A.F. Bobick. 2000. Recognition of visual activities and interactions by stochastic parsing. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 8 (2000), 852–872.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, Vol. 2. IEEE, 2169–2178.

R. Mann, A. Jepson, and J.M. Siskind. 1997. The computational perception of scene dynamics. *Computer Vision and Image Understanding* 65, 2 (1997), 113–128.

E. T. Mueller. 2006. *Commonsense Reasoning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

J. Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press, New York, NY, USA.

M. Pei, Y. Jia, and S.-C. Zhu. 2011. Parsing video events with goal inference and intent prediction. In *ICCV*. 487–494.

K. Prabhakar, S. Oh, P. Wang, G.D. Abowd, and J.M. Rehg. 2010. Temporal causality for the analysis of visual events. In *CVPR*.

M. Richardson and P. Domingos. 2006. Markov logic networks. *Machine Learning* 62, 1 (2006), 107–136.

D.B. Rubin. 2005. Causal inference using potential outcomes. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.

Rebecca Saxe and Susan Carey. 2006. The perception of causality in infancy. *Acta psychologica* 123, 1 (2006), 144–165.

R. Saxe, JB Tenenbaum, and S. Carey. 2005. Secret Agents Inferences About Hidden Causes by 10-and 12-Month-Old Infants. *Psychological Science* 16, 12 (2005), 995–1001.

R. Scheines, H. Hoijtink, and A. Boomsma. 1999. Bayesian estimation and testing of structural equation models. *Psychometrika* 64, 1 (1999), 37–52.

P. Spirtes, C. Glymour, and R. Scheines. 2000. *Causation, prediction, and search*. Vol. 81. MIT press.

S. Tran and L. Davis. 2008. Event modeling and recognition using markov logic networks. *ECCV* (2008), 610–623.

Z. Tu, X. Chen, A.L. Yuille, and S.-C. Zhu. 2005. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision* 63, 2 (2005), 113–140.

P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu. 2013. Modeling 4d human-object interactions for event and object recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 3272–3279.

S.-C. Zhu and D. Mumford. 2006. *A Stochastic Grammar of Images*. Now Publishers Inc., Hanover, MA, USA.

S.-C. Zhu, Y.N. Wu, and D. Mumford. 1997. Minimax entropy principle and its application to texture modeling. *Neural Computation* 9, 8 (1997), 1627–1660.