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

Amy Fire (amy.fire@ucla.edu)
Song-Chun Zhu (sczhu@stat.ucla.edu)
Center for Vision, Cognition, Learning, and Art
University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

For both human and machine learners, it is a challenge to make high-level sense of observations by identifying causes, effects, and their connections. Once these connections are learned, the knowledge can be used to infer causes and effects where visual data might be partially hidden or ambiguous. In this paper, we present a Bayesian grammar model for human-perceived causal relationships that is learnable from video. Two experiments investigate high-level causal induction from low-level visual cues. In the first experiment, we show that a computer can apply known heuristics used for causal induction by humans to learn perceptual causal relationships. In the second experiment, we show that our learned model can represent humans' performance in reasoning about hidden effects in video, even when the computer initially misdetects those effects.

Keywords: Perceptual causality; causal induction; statistical models.

Introduction

A man approaches a closed door. He reaches out to grasp the handle and then stands there. Is it locked? Does he not have the key? He knocks and waits, but the door remains closed. Is there no one on the other side to open it?

Watching these events unfold, humans can readily answer these questions based on their causal knowledge. One way humans can learn causal relationships is through daily observation by internally measuring co-occurrence of events (Griffiths & Tenenbaum, 2005). Research suggests that humans use a few heuristics to determine whether a co-occurrence is causal, including:

- whether the temporal lag between cause and effect is short, and the cause precedes the effect (Carey, 2009) and
- whether agent actions are responsible for causes (Saxe, Tenenbaum, & Carey, 2005).

However, learning from daily observation is limited: many actions and effects are hidden. Our prior knowledge about causal relationships between actions and effects allows us to fill in information about the events in the scene.

Some current models represent knowledge with Bayesian networks, e.g., (Griffiths & Tenenbaum, 2005). These models, however, are disjoint from the low-level visual data that people observe. Instead, models are built using high-level annotations. In reality, agents build knowledge by observing low-level visual data, and models need to be able to deal with uncertainty in observation.

Although Bayesian networks are commonly used to represent causality (Pearl, 2009), grammar models have the expressive power to represent a greater breadth of possibilities than a single instance of a Bayesian network (Griffiths & Tenenbaum, 2007). Grammar models allow for multiple configurations and high-level structures, making them more suitable for applications grounded on visual cues; Bayesian networks lack the representative power needed for this.

Grammar models are represented graphically in the And-Or Graph (AOG). In the AOG, Or-nodes represent the multiple alternatives, and And-nodes represent hierarchical decompositions. The AOG naturally lends itself to represent causation where multiple alternative causes can lead to an effect, and each cause is composed of conditions necessary for the effect.

In this paper, we introduce a grammar model for representing causal relationships between actions and object-status changes, the Causal And-Or Graph (C-AOG). We describe methods for learning the model by using co-occurrence to identify potential causal relationships between events and applying the heuristics listed above to those potential relationships. In two experiments, we investigate how the model matches human perceptions of causality. Experiment 1 uses input typical of computer vision detection systems to investigate learning the C-AOG and human perceptions of causality. Experiment 2 demonstrates that the C-AOG models human judgments on imputing hidden variables from video.

A Grammar Model for Causality

In this section, we introduce the Causal And-Or Graph for causal reasoning, which ties agent actions to fluents.

Fluents and Actions

Specifically defining those object statuses that vary over time, the term *fluents* comes from the commonsense-reasoning literature (Mueller, 2006). Relevant here are two kinds of fluents that intentional agents can change: object fluents (e.g., a light can be on or off) and fluents of the mind (e.g., an agent can be thirsty or not thirsty).

The values of these fluents change as a result of agent actions and also trigger rational agents to take action. A lack of change-inducing action (also known as the *inertial action*) causes the fluent to maintain its value; for example, a door that is closed will remain closed until some action changes it. In this work, fluents are modeled discriminatively.

Actions (A_i) are modeled using the Temporal And-Or Graph (T-AOG), a grammar model for actions (Pei, Jia, & Zhu, 2011). In the T-AOG, And-nodes group the necessary ways for an action to be performed that allow detection of the action (e.g., object/agent spatial relations, agent poses, scene contexts, and temporal relationships), and Or-nodes provide

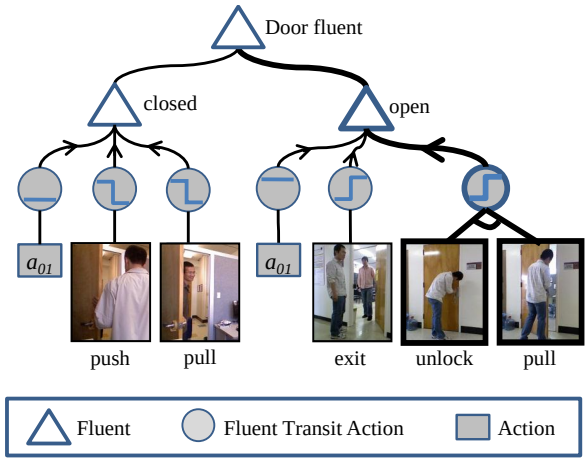


Figure 1: A C-AOG for door status as learned in Experiment 1. The value of the top-level fluent is a consequence its children. The fluent transit action nodes indicate the kind of change that occurs in the fluent: step functions for change, flat lines for non-change (or inertial action). Action a_0 is the inertial action (a lack of state-changing action). Arcs connect children of And-nodes. It should be noted that each photo represents a further set of child And-nodes from the Temporal And-Or Graph (not shown). Thickened lines indicate selections on the Or-nodes that provide a single parse graph.

the alternative methods of performing the action. While hidden Markov models and dynamic Bayesian networks have also been used for action detection from video, the grammar is necessary as it allows representation of high-level structures and multiple configurations.

Our experiments are conducted using a pre-selected set of actions and fluents common to office, hallway, and elevator scenes. Such scenes (and events therein) might be of interest for surveillance, for example.

The Causal And-Or Graph

The Causal And-Or Graph (C-AOG) is a graphical representation for the grammar of causality. The top levels of one C-AOG learned in Experiment 1 are shown in Figure 1.

In the C-AOG, Or-nodes represent the alternative means of causation (e.g., a monitor, through the computer, can be turned on by someone using a mouse or a keyboard). Arrows point from these causing actions to their fluent effects.

Each And-node is formed from the set of multiple conditions for the action, including its sub-actions. The action nodes in a C-AOG may be inertial actions (resulting in no change); unexplained instances of the fluent are also pooled under the inertial action.

A selection on the Or-nodes is called a parse graph, denoted *pg* (such as the paths shown by thicker lines in Figure 1). It provides a causal interpretation of each fluent's particular value at a given time, answering "why" the fluent has that particular value.

Probability on the C-AOG

The probability model over the parse graphs in the C-AOG incorporates the detection probabilities of actions and fluents in a Bayesian manner. In particular, given the video I,

$$\underbrace{P(pg_C|I)}_{\text{posterior}} = \underbrace{P(A_1,\ldots A_n|I)P(\Delta F_1,\ldots,\Delta F_m|I)}_{\text{likelihood}} \prod_{v\in V_C^{\text{Or}}} \underbrace{P(w(v))}_{\text{prior}}. \tag{1}$$

The likelihood term is the detection probability for the included actions/fluents, and considers actions and fluents independently. V_C^{Or} is the set of included Or-nodes in the causal explanation, and $w(v)$ returns the selected Or-branch. The prior term gives the switch probability on the Or-nodes for the alternative causes and is learned by maximum likelihood estimation.

Learning the C-AOG

To learn the C-AOG, potential causal relationships are found by restricting the set of all possible fluent/action interactions with the set of heuristics listed at the beginning. Actions and fluents from all levels of their respective hierarchies are considered.

A joint model is iteratively built up from the initial probability distribution over actions and fluent changes, incorporating a new causal relationship each iteration. In an iteration, the contingency table of each action-fluent pair $(A_i, \Delta F_j)$, e.g., Table 1, is examined. The best causal relationship is determined by maximizing the information gain (IG), which is the Kullback-Leibler divergence (KL) (Kullback & Leibler, 1951) between the full contingency table of Table 1 and the expected contingency table predicted by the model in the current iteration (similar to work on texture modeling (Zhu, Wu, & Mumford, 1997)). In particular, in a single iteration, causal relation cr^* is added to the model where

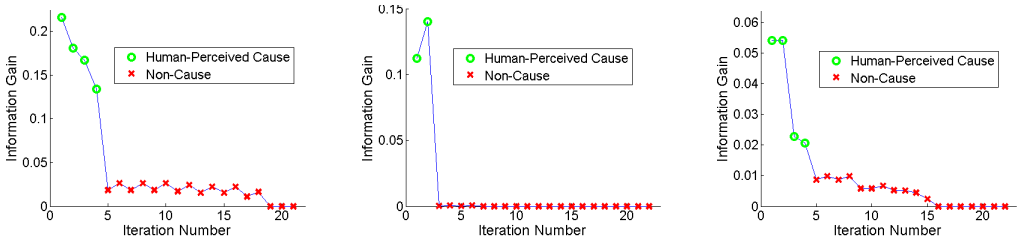
$$cr^* = \underset{cr}{\text{argmax}}\,\text{IG} = \underset{cr}{\text{argmax}}\,\text{KL}(\mathbf{f}||\mathbf{h}), \tag{2}$$

$\mathbf{f} = (f_0, f_1, f_2, f_3)$, and $\mathbf{h}$ is the analogous quantity from the current iteration's model. The causal relationships with highest information gains are deemed most significant and are collected into the C-AOG.

Table 1: Contingency table of relative frequencies.

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

Our learning method integrates with existing action and fluent detection systems, creating a unified framework for the spatial, temporal, and causal domains. Further, our method is more computationally feasible for large networks of causal connections than Bayesian learning frameworks are (with their prior distributions over graph structures). Traditional causal induction as done by constraint satisfaction (Pearl,



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

Figure 2: Information gains for causal relations in the order pursued, separated by fluent. Green circles label causes.

2009) or Bayesian formulations (Heckerman, 1995) is intractable to ground on vision sensors. Models such as causal support (Griffiths & Tenenbaum, 2005) learn a new, larger model each iteration, and the number of possible models grows exponentially. In contrast, the number of computations to learn our model is constant each iteration.

Experiment 1: Learning Causality

In this experiment, we test the model's ability to learn human-perceived causal relationships. For testing the algorithm, the ground truth is established by linking known causing actions to their fluent effects.

Video Data Used

To test learning the C-AOG, videos were collected with a Microsoft Kinect, recording the color and depth images simultaneously. The scenes collected include multiple doorways, an elevator, and an office. Figure 1 shows some screenshots of the videos. The entire video collection lasts about 120 minutes, and contains 21 pre-specified action categories. There are 8 to 20 (sometimes simultaneous) instances of each action category.

In this experiment, we first use perfect action and fluent detections to demonstrate learning. We compare these results to those obtained with noisy detections (with varying levels of accuracy), such as would be output from the action and fluent detection system.

Results and Discussion

Multiple Fluents Figure 2 shows plots of information gains for causal relations in the order pursued, separated by fluent. Causes are added to the model before non-causes with clear cutoffs for the door and light fluents. The cutoff between cause and non-cause is obscure for the computer monitor fluent because the model only acquired partial causal information (the preconditions of power and computer status are hidden).

Noisy Data Randomly flipping action detections leads to the curves shown in Figure 3. 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.

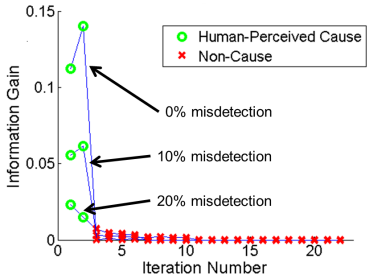
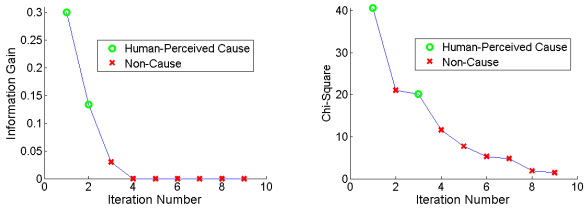


Figure 3: Information gains for causal relationships in the order pursued for the light fluent.

Hierarchical Action Selection and χ^2 Where compound actions (e.g., in the doorway scene, unlocking with a key or entering a code, followed by pushing/pulling the door) are required for the effect, the causing actions may be located within varying levels of the action hierarchy.

For actions hierarchically related to each other in the Temporal AOG, our model incorporates their dependences, minimizing the chance that related actions are selected as causes. Figure 4 shows that Hellinger's χ^2 measure (a χ^2 that is less sensitive to low expected values in a contingency table (Ferguson, 1996)) fails to identify the correct causes, unable to account for dependence.



(a) Our Method (b) Hellinger's χ^2

Figure 4: Pursuit order for hierarchical causes.

Long Delay, Causal Power, and ΔP Under the power PC theory (Cheng, 1997), perceptual causality is calculated as:

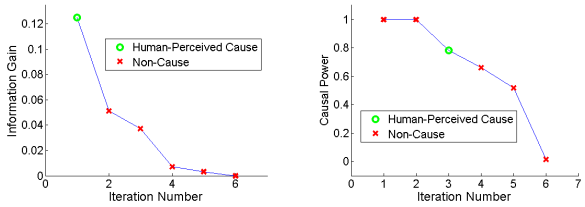
$$\text{causal power} = \frac{\Delta P}{P(\text{effect}|\text{not cause})} \tag{3}$$

Figure 5: Sample of human judgment key frames.

where ΔP (Allan, 1980) is given by:

$$\Delta P = P(\text{effect}|\text{cause}) - P(\text{effect}|\text{not cause}). \quad (4)$$

For an elevator, the only detectable causing action for the door opening is pushing the elevator call button. In this example, our model outperforms causal power as shown in Figure 6. ΔP performs similarly to causal power.



(a) Our Method (b) Causal Power

Figure 6: Pursuit order for the elevator scene.

The failure of causal power and ΔP originates when an observed event (e.g., walking away) coincidentally always occurs with the true cause (e.g., pushing the elevator call button) and the true cause is not perfectly detected. Both measures favor 100% correlation, despite how rarely it occurred in the video. The learning method presented here incorporates the frequency that the relationship is observed by examining the full contingency table.

Further Discussion

Results match exactly with human perceptions of the causal connections between actions and fluent changes, showing that the C-AOG is learnable from co-occurrence and the heuristics listed in the beginning (short temporal lag and agent actions cause fluent changes).

Our results are limited to the action and fluent categories that are pre-specified, despite the fact that many potentially confounding actions were included. Those quantities must be specified in advance so that appropriate detectors can be trained. It is possible, however, that different people would produce different bottom-level actions and fluents.

Experiment 2: Inference Experiment

In this experiment, our model is validated against humans in the long-term reasoning task of inferring hidden fluent values.

The Stimuli

Approximately 20 minutes of video data was captured using a Kinect in two scenes: a hallway and an office. Table 2 contains a summary of the fluents contained in the video, as well as the values each fluent can take. While many of these fluents are ordinarily viewable, they are ambiguous in the video (e.g., light status (ambient light may be from a window or a light) or water stream (resolution is not high enough to see it) in Figure 5).

Through a website, volunteer participants ($N = 15$) were shown the test video which paused at preset frames, e.g., those shown in Figure 5. Query points surround either a change in a fluent or a causing action. At each key frame, the participant was asked to assign a total of 100 points to all possible values of each fluent, according to his/her own recognition and reasoning for the events. Assignment of the points corresponded to the subjective probabilities of the fluent values. Each participant was allowed to revise previous judgments with information derived from subsequent frames.

Reference Estimates

We compare the human responses to predicted fluent values by a baseline random noise model and by the C-AOG.

Baseline Estimate (Random Noise). For a baseline estimate, the hidden fluents were randomly assigned uniformly, without using any detection or causal information (e.g., 50% for LIGHT ON and 50% for OFF). The baseline estimate provides a discriminative reference against which we can see how well our model approximates human judgments.

Computer Estimate (The C-AOG). Detectable actions and fluent changes are first extracted from the videos and used

Table 2: List of fluents considered.

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

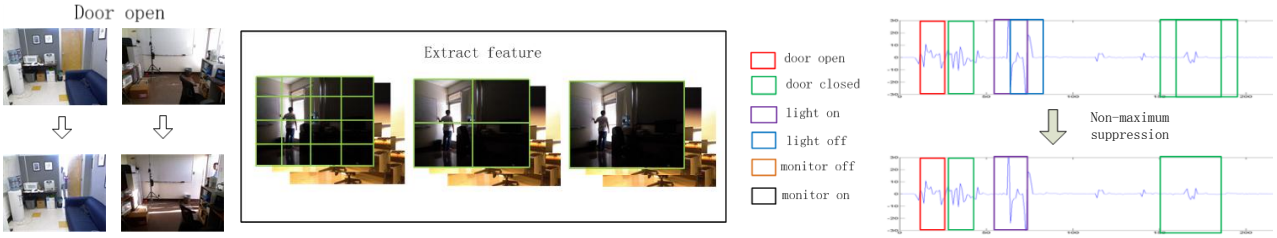


Figure 7: Sample screenshots for noisy data.

as inputs to the C-AOG model.

The action grammar is pre-specified. Actions are manually segmented, and then poses captured by the Kinect camera are clustered. Temporal parsing transforms the clustered poses into hierarchically-labeled instances from the T-AOG. The maximum probability action detections are used as input.

Fluent changes are detected from the video with the GentleBoost algorithm (Friedman, Hastie, & Tibshirani, 2000) on features extracted as shown in Figure 7. Non-maximum suppression provides the final detections of fluent changes.

These action and fluent detections (and their probabilities) are then processed with potential causal explanations under the C-AOG (by maximizing the posterior probability of Equation 1). The best-performing consistent causal description over the course of the video is then returned through the Viterbi algorithm (Forney Jr, 1973). Hidden fluents are imputed from this result.

Results and Discussion

To visualize the results, human, computer, and baseline estimates are reduced to two dimensions using multi-dimensional scaling (MDS) according to the total variation distance between estimates, and plotted in Figure 8.

In the hallway dataset, both fluent and action detections contribute to the causal inference of hidden fluents. The computer performance is very similar to human performance as shown in Figure 8(a). The baseline is far from the cluster of computer and human estimates.

The office dataset only contains detections of actions; all fluents are hidden. The computer's performance is still an improvement over the baseline towards human-level performance, as shown in Figure 8(b).

Misinformation: Correcting Spatio-Temporal Detections In the hallway dataset, multiple changes in the light fluent were detected, yet no causing action was detected, presenting a common situation in vision—detections are usually imperfect. The C-AOG corrects these errors by balancing the maintenance of detections with the consistency of causal explanations. Figure 9 shows typical candidates of the results sorted in order of probability.

The C-AOG result was consistent with human judgments. Humans selected a single value for the light fluent for the duration of the video, but some selected ON while others chose OFF. This reinforces the need to have a probabilistic model

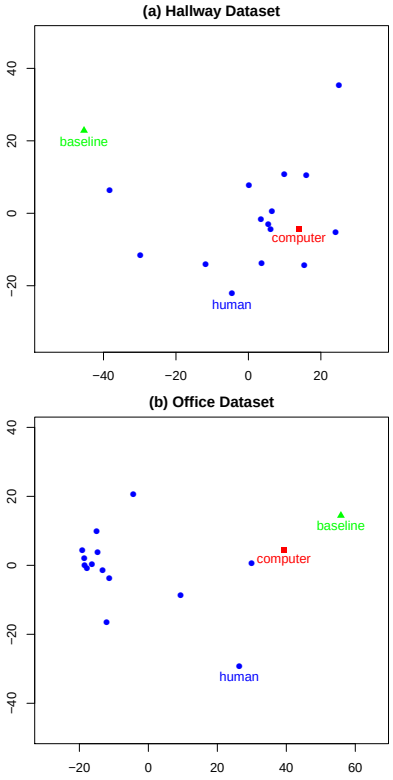


Figure 8: MDS plots of fluent value estimates. Blue dots: human estimates. Red squares: estimates using the C-AOG. Green triangles: baseline estimates. See Further Discussion for notes on the human variability.

capable of maintaining multiple interpretations; the C-AOG result included both solutions.

Further Discussion

Even though the set of possible fluent values was provided to participants (significantly narrowing their available judgments), the MDS plots show wide variation in human responses. This is due to many factors. First, some participants initialized fluent values differently (e.g., light ON versus OFF in Figure 5), resulting in a large total variation distance. Also, some participants were more cautious than others, recording judgments close to 50/50 where others took an all-or-nothing approach to assigning judgments.

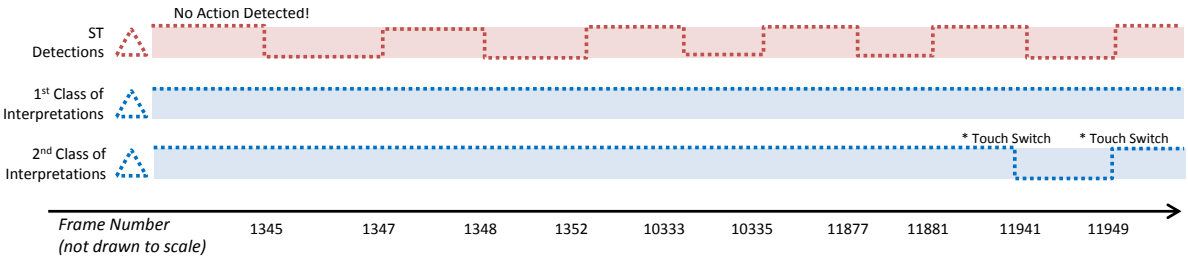


Figure 9: Given action and fluent detections that move the light fluent between ON and OFF without a causing action, the C-AOG prefers this to be explained by incorrect detections of the light fluent. The second most probable class of explanations is that two of the changes had causing actions that were missed by the detection.

As evidenced by the C-AOG's weaker performance, the office dataset was particularly challenging. Action detections were poor and no fluent detections were available to identify conflicts, leaving the system heavily dependent on those incorrect action detections. Despite this disadvantage, the C-AOG still provided enough reasoning capability to outperform the baseline. This example underscores the importance of good vision-detection systems.

Conclusions and Next Steps

In this paper, we have presented a probabilistic graphical grammar model to match human perception of causal relationships between actions and fluent changes, the Causal And-Or Graph (C-AOG).

Experiment 1 showed that the C-AOG of everyday activities can be learned, matching human perceptions of causal relationships. These causal relationships are even learnable amid noise, such as would be present in detection systems. Further, experiment 1 showed that our method models human judgments better than causal power and ΔP.

Experiment 2 showed that the C-AOG can be used as a model of human perception grounded on video to impute values for hidden fluents. This experiment captures the inherent variability of human estimations when confronted with video, and highlights the need for a model that can probabilistically incorporate causality and vision.

One current limitation of the C-AOG is that, if a situation is unexplained, all possible parse graphs are assigned a low probability. In future work, we plan to investigate how adaptive learning can be used to incorporate new instances of fluents into the C-AOG.

Acknowledgments

This work is supported by ONR MURI grant N00014-10-1-0933.

References

Allan, L. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149.

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

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological review*, *104*(2), 367.

Ferguson, T. (1996). *A course in large sample theory: Texts in statistical science* (Vol. 38). Chapman & Hall/CRC.

Forney Jr, G. (1973). The viterbi algorithm. *Proceedings of the IEEE*, *61*(3), 268–278.

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

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

Griffiths, T., & Tenenbaum, J. (2007). Two proposals for causal grammars. *Causal learning: Psychology, philosophy, and computation*, 323–345.

Heckerman, D. (1995). A bayesian approach to learning causal networks. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence* (pp. 285–295).

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Mueller, E. T. (2006). *Commonsense reasoning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

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

Pei, M., Jia, Y., & Zhu, S.-C. (2011). Parsing video events with goal inference and intent prediction. In *Computer vision (iccv), 2011 ieee international conference on* (pp. 487–494).

Saxe, R., Tenenbaum, J., & Carey, S. (2005). Secret agents inferences about hidden causes by 10-and 12-month-old infants. *Psychological Science*, *16*(12), 995–1001.

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