Learning Perceptual Causality from Video

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

Abstract

Computer vision and artificial intelligence research has long danced around the subject of causality: vision researchers use causal relationships to aid action detection, and AI researchers propose methods for causal induction independent of video sensors. In this paper, we argue that learning perceptual causality from video is a necessary step for understanding scenes in video. We explain how current object and action detection is suffering without causality, and we explain how current causality research is suffering without grounding on raw sensors. We then go on to describe one plausible solution for grounding perceptual causality on raw sensors.

Applying causal knowledge to vision research provides a much deeper level of understanding than considering actions and objects independently. Causal understanding enables joint spatial-temporal-causal inference (allowing causal information to connect spatial and temporal domains). With joint inference, it becomes possible to infer misdetected and hidden objects and actions, along with an agent's state of mind, intents, and goals.

To understand the depth of how causality can be used, imagine a scene where an agent flips a switch and a light turns on. The light switch itself is hard to detect. However, causality bridges the gap between the spatial and temporal detections of the light turning from off to on and the agent approaching the wall (Figure 1(b-c)). We can then use causality to infer with high probability that the object on the wall is the light switch and that the agent was performing the action of turning the light on, even when the switch is occluded, not otherwise detectable, or misdetected.

Further, we can assume the agent turned the light on because he desired it thus. Knowing that the light was initially off, an observer could infer that an agent would most likely turn it on upon entering the room.

Under the goal-driven stance taken by cognitive science researchers (Csibra and Gergely 2007), the agent's action occurred because he wanted to change some facet of the world. Both the agent's intent and the previous state of the world are preconditions for the causing action, and both can be inferred once the causal connections are understood.

Watching the agent next move to a drinking fountain and take a drink, an observer can infer that the agent was thirsty because thirst is a common trigger for drinking. After the agent moves from the fountain, the observer can infer that the agent is most likely sated, having satisfied his thirst.

By connecting preconditions, trigger conditions, actions, and effects over time, one can infer the most probable consistent explanation (Figure 1(d)), filling in missing values over the course of the video.

Causality completes many of the missing detections from the scene by connecting states of the world and agent actions. Most humans use a simplified causal model when answering questions. When asked what caused the light to turn on in the example presented, humans will identify the agent's action alone—ignoring all the other necessary conditions for the effect such as working electrical power and the switch being connected to the light (Mackie 1965).

The learning of causality is largely missing from the vision literature. We argue that this learning can be accomplished *from video* by building on current research in cognitive science, using the explanations that humans volunteer as a basis. Specifically, it is possible to:

1. Learn perceptual causality, acquiring the knowledge of causal relationships in the world as humans do.
2. Infer instances of perceptual causality (jointly with spatial and temporal inference), using the learned causal knowledge to identify instances in video.

Perceptual Causality

As humans observe their world, they form conclusions about causal relationships, linking states of the world to perceived causing conditions. Cognitive scientists recognize that even infants are equipped with a notion of *perceptual causality*, able to draw causal conclusions from observations based on temporal spacing and an innate understanding of agency (Carey 2009). It is this type of causality that we argue is learnable from video.

A starting place for vision research to acquire perceptual causality is to examine the connection between actions and fluents. A *fluent* is defined in the commonsense-reasoning literature as an object status that specifically varies over time (Mueller 2006). For example, a light's fluent takes the values "on" and "off" over time as the light turns on and off.

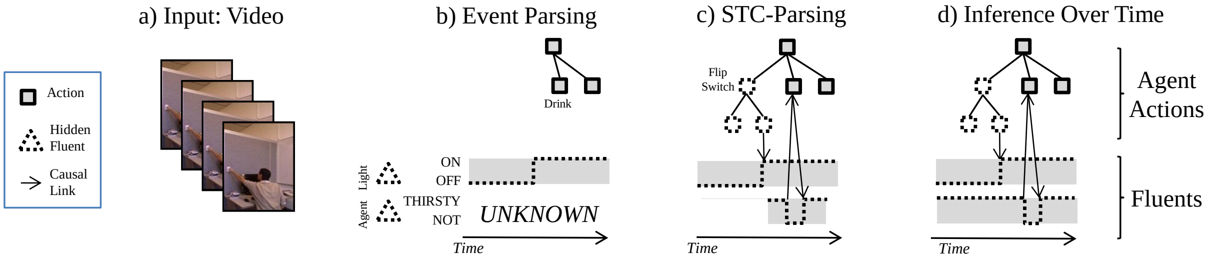

Figure 1: An example of the inference that is possible by using causality. a) Sample video input. b) A possible parsing of the video attainable using only the spatial and temporal domains, without causal knowledge. Actions are parsed from a hierarchy, using the Spatio-Temporal And-Or Graph (Pei, Jia, and Zhu 2011). c) A possible parsing attainable with causal knowledge. d) The full explanation (filling in the agent's thirst) over the course of the video is attainable with causal knowledge over time.

For a typical, functioning light, it is an agent's actions that dictate the light's fluent value through a causal connection. Humans perceive the agent flipping the switch as causing the light to go on or off, even though there are many necessary conditions as well.

The importance that humans place on agentive action in determining causes is one key in acquiring perceptual causality. Once perceptual causality is learned from video, it can be used for the many inference tasks discussed in the beginning.

Vision Research and Causal Knowledge

A complete integration of causal knowledge with detection systems for objects and actions is missing from the vision literature. Current works tend to study object fluents independently of causing actions.

Spatial co-occurrence in a single frame enables detection of objects and object fluent values in a scene from the raw sensor data of an image (Tu et al. 2005). However, as these rely only on single images (not video with actions being performed), there is necessarily a lack of causal understanding integrating actions and fluents.

Some researchers have used causal relationships together with temporal co-occurrence for action recognition (Albanese et al. 2010), focusing on high-level descriptions of the video, but these works do not attempt to learn causality. Following the causal connections granted by physics, Newtonian mechanics have also been used to distinguish actions (Mann, Jepson, and Siskind 1997).

Recently, Prabhakar et al. used causal measures to learn patterns of repeated actions, identifying visual words as point processes (2010). Their method relates the low-level visual events using Granger causality, allowing the unsupervised identification of independent sets of these events. However, their method does not learn, nor infer, causality.

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

None of these approaches to temporal analysis begin to approach causality in a way that allows it to be learned from video. A viable manner of learning causality from video data will integrate with both spatial and temporal learning strategies at the pixel level to provide a coherent solution.

Causality and Video Data

Learning causality in artificial intelligence, on the other hand, usually amounts to traditional causal induction as done by constraint satisfaction (Pearl 2009) or Bayesian formulations (Heckerman 1995). These methods are intractable to ground on vision sensors. Even using these systems atop mid-level visual words is computationally infeasible when considering the vast domain of observable causal relations.

Causal inference in commonsense reasoning is usually solved by first-order logic (Mueller 2006). However, these deductive methods do not allow for probabilistic solutions, which are needed in computer vision to allow for ambiguity given that detections are often unreliable.

Markov logic networks (Richardson and Domingos 2006) relax the strictness of first-order logic by placing a Markov random field atop the first-order logic. They were applied to the task of action detection (Tran and Davis 2008), but the knowledge base formulas used for the logic were not learned. Further, Markov logic networks are solved with Gibbs samplers and are intractable for general inference from low-level vision sensors.

The theories for learning and inferring causality that have been developed in artificial intelligence are insufficient for the task of learning from video. Even though perceptual causality lacks the accuracy of causal induction, it can still provide valuable information.

Joining Vision and Causality

Humans learn perceptual causality through daily observation by internally measuring co-occurrence of events and effects (Griffiths and Tenenbaum 2005). Research of infants shows that, in pursuing causal knowledge, this co-occurrence is restricted to events where the temporal lag between cause and effect is short (Carey 2009), cause precedes effect (Carey 2009), and agentive actions are causes (Saxe, Tenenbaum, and Carey 2005).

Analogous observation for a computer comes through video, and to begin learning perceptual causality, the computer must examine this co-occurrence, similarly restricted.

Beginning with a vision system that detects fluents and actions from video, this method can learn causality from video in an unsupervised manner. Further, by using the same measure for co-occurrence to learn objects and actions from low-level sensors as used for learning perceptual causality, we can provide a principled approach to learning (Fire and Zhu 2013).

To represent causal knowledge, a grammar model is needed (Bayesian networks lack their expressive power) (Griffiths and Tenenbaum 2007). Grammar models are embodied graphically in the And-Or Graph. The Causal And-Or Graph (pictured in Figure 2) represents a grammar of causality: And-nodes group necessary conditions, and Or-nodes provide alternate causes. Arrows point from causes to effects. The Causal And-Or graph creates another layer of hierarchy atop spatial and temporal grammar models, which are grounded on raw sensors.

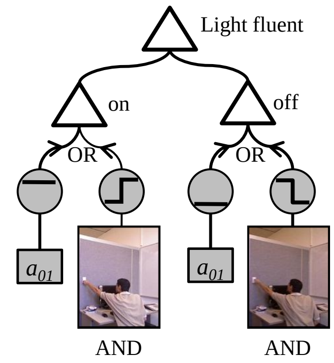

Figure 2: A simple Causal And-Or Graph for the light status. The light can be "on" because an agent flipped the switch (changing the light from "off" to "on"), or because the light was already "on" and no change-inducing action occurred (shown with a_{01}). On the lowest level pictured here, the photos for the actions represent And-nodes comprising of relations for action detection. Similarly for the light "off".

Discussion

In vision research, one of the main goals is for the computer to understand the images and videos. Complete understanding is important for applications such as situationally aware robots and intelligent video-surveillance systems. Causal knowledge is important to that understanding.

The task of acquiring causal knowledge is a challenging one. Detection of causal relationships relies on the accurate detection of both causes and effects. Hand-labeling static objects in a single frame of the video can greatly improve detection. Further, specifying dictionaries of bottom-level fluents and actions can simplify the search space.

As a starting point, we propose to examine the perceptual causal link between actions and fluents by examining co-occurrence subject to "commonsense" heuristics. The perceptual causal knowledge acquired enhances the computer's understanding of video, adding an explanation of why fluents change (because an agent's action changed them) and why actions are most likely performed (to change fluents under the goal-driven view of the human mind).

Acknowledgements

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

References

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

Brand, M. 1997. The "inverse hollywood problem": From video to scripts and storyboards via causal analysis. In *Proceedings of the NCAI*, 132–137.

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

Csibra, G., and Gergely, G. 2007. Obsessed with goals: Functions and mechanisms of teleological interpretation of actions in humans. *Acta psychologica* 124(1):60–78.

Fire, A., and Zhu, S.-C. 2013. Learning and inferring causality from video. *Pre-Print*. Department of Statistics, UCLA.

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

Griffiths, T., and 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 UAI*, 285–295. Morgan Kaufmann Publishers Inc.

Mackie, J. 1965. Causes and conditions. *American philosophical quarterly* 2(4):245–264.

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

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

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

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

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

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

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

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

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