CAUSAL DISCOVERY IN VIDEO VIA DNNS

Haoran Liu, Qijun Yang, Elinor Cheng
Paper under double-blind review

ABSTRACT

In our research, we enhance object detection and the analysis of causal links in video data using Neural Granger models. By applying cMLP and cLSTM, we delve into the temporal connections between objects to reveal underlying causal dynamics. This integration of sophisticated object detection and tracking establishes a solid foundation for identifying causality in video sequences, contributing significantly to the field of dynamic system analysis. Our methodology offers a novel perspective on video-based causality, opening avenues for deeper insights into the interactions within complex systems.

1 INTRODUCTION

Video analysis has emerged as a crucial domain in computer vision, encompassing a wide range of applications, from surveillance and autonomous vehicles to human-computer interaction. The ability to extract meaningful insights from video data is of paramount importance in understanding complex dynamic systems. One fundamental aspect of this analysis is the identification of causal relationships among objects or subjects within videos, as this can provide valuable insights into the underlying dynamics of the observed scene Fire & Zhu (2015).

Causal discovery in video data is a challenging and intriguing problem. Traditional methods for causal inference have primarily focused on static data, making them ill-suited for the dynamic nature of videos. In recent years, the advent of deep neural networks (DNNs) has opened up new avenues for addressing this challenge. DNNs have demonstrated remarkable capabilities in various computer vision tasks, and their potential to unearth causal relationships within video data has garnered significant attention.

In this paper, we present a novel approach for causal discovery in video data using component-wise Multilayer Perceptrons (cMLP) and component-wise Long Short-Term Memory networks (cLSTM) Tank et al. (2021). Building upon the concept of Granger causality, we leverage the temporal dependencies within video sequences to unveil causal relationships among objects or subjects. Our methodology combines advanced object detection techniques, including YOLO8 Redmon et al. (2016) and Faster R-CNN Girshick (2015), with object tracking via DeepSORT Wojke & Bewley (2018), creating a comprehensive framework for causal discovery in videos.

2 METHODOLOGY

Our project is structured into three distinct phases to achieve our goals:

PHASE 1: DATASET SELECTION AND AUGMENTATION

We begin by selecting a video dataset for object detection, focusing on scenarios with multiple interacting objects, some of which are obstructed. This dataset will then be augmented to enhance our algorithms' ability to manage obstructed objects.

PHASE 2: VIDEO-BASED OBJECT DETECTION

The second phase involves implementing two algorithms for video-based object detection, tailored to address missing or obstructed data. These algorithms will identify and track objects within the video, forming the basis of our object detection system.

PHASE 3: GRANGER CAUSALITY ESTIMATION WITH NEURAL NETWORKS

Finally, we will develop one or two algorithms for Granger causality estimation. These will analyze time-series data from the video to identify causal relationships between objects.

Through these phases, we aim to create a comprehensive system capable of effective object detection and causal analysis in dynamic video environments, thereby addressing the complex objectives of our project.

3 IMPLEMENTATION

3.1 DATA AUGMENTATION

Our dataset selection revealed that video frames often don't show all key objects. To tackle this and teach our model to recognize partially visible objects, we adopted a data augmentation approach. We extract individual frames and **randomly mask** players in them by setting their bounding box pixels to zero, effectively making them invisible. This strategy introduces variety in our training data, challenging our object detection model to identify and label objects even when obscured. This augmentation prepares our model for real-world scenarios, improving its robustness and accuracy in situations where objects might be temporarily hidden.

3.2 DETECTION ALGORITHM

3.2.1 FASTER R-CNN

Faster R-CNN Girshick (2015), an advanced deep learning framework, excels in object detection and localization. Improving upon R-CNN, it combines convolutional neural networks (CNNs) with **region proposal networks (RPNs)** for precise object identification in images and videos. In our project, we utilize the pre-trained **ResNet50** He et al. (2016) for feature extraction from video frames. The innovation lies in retraining the RPN in the Faster R-CNN, allowing quick generation of bounding boxes and object labeling. For instance, in a basketball game video, Faster R-CNN enables accurate tracking and identification of players, including jersey numbers, demonstrating the synergy between ResNet50's layers and the optimized RPN.

3.2.2 YOLOv8

The eighth version in the "You Only Look Once" series, is a breakthrough in object detection algorithms, renowned for its speed and accuracy Jocher et al. (2023). Tailored for real-time analysis, it processes images in a single pass, making it ideal for tasks requiring swift and precise object identification. This version

surpasses its predecessors with significant improvements in deep learning, enhancing its efficiency and detection capabilities.

In our project, YOLOv8 is employed to identify and track basketball players in a video. The goal is to accurately recognize players and their movements during a game, a task complicated by the sport's fast pace and the need to monitor multiple players simultaneously. YOLOv8 has been fine-tuned to distinguish players from other elements in the video, processing each frame to provide real-time data on their positions and actions. Its advanced features, including the ability to manage varying object sizes and fast processing speeds, make YOLOv8 highly effective for this demanding application.

3.3 TRACKING ALGORITHM

3.3.1 DEEPSORT

We incorporate DeepSORT on top of a complex detection model like YOLO allows for consistent labeling of the same players throughout a video sequence, significantly improving tracking robustness. DeepSORT extends the SORT (Simple Online and Realtime Tracking)Wojke et al. (2017) algorithm by integrating deep learning features for improved tracking performance. The process for each frame is as follows:

1. The detector obtains bounding boxes and generates detections.
2. The Kalman filter predicts the new state of each track.
3. The Hungarian algorithm matches the predicted tracks with the current frame detections using cascade matching and IOU matching.
4. The Kalman filter updates the tracks with the matched detections.

For instance, at Frame 0, the detector identifies three detections. As there are no existing tracks, these detections are initialized as new tracks. Moving to Frame 1, the detector again identifies three detections. The existing tracks from Frame 0 are predicted to the new locations using the Kalman filter, after which the Hungarian algorithm matches these new predicted tracks with the detections. Each match consisting of a (track, detection) pair is used to update the corresponding track.

The Hungarian Algorithm assigns consistent labels based on the previous frame to the following frame. It operates the distance matrix and finds the best match between frames. The Kalman filter is utilized for predicting the future locations of objects from their current states. It operates on the principle of estimation and correction, using a series of measurements observed over time, containing statistical noise and other inaccuracies. The main equations for the Kalman filter in the context of tracking are the state prediction and covariance prediction equations:

$$\begin{aligned} x' &= Fx \\ P' &= FPF^T + Q \end{aligned} \tag{1}$$

Here, x represents the state vector at time $t-1$, x' is the predicted state vector at time t, F is the state transition matrix which models the dynamics of the system, P is the covariance matrix of the track at time $t-1$, and Q is the process noise covariance matrix which accounts for the uncertainty in the model.

3.4 GRANGER CAUSALITY

Granger causality is a method to investigate the causality between two time series. It is based on the premise that if a time series X Granger-causes another time series Y, then past values of X should contain information that helps predict Y beyond the information contained in past values of Y alone.

A typical Granger causality model can be represented as a vector autoregression, where the current value of Y_t is regressed on its own past values and the past values of X_t:

$$Y_t \approx \sum_{j=1}^{d} a_j \cdot Y_{t-j} + \sum_{j=1}^{d} b_j \cdot X_{t-j} \tag{2}$$

Alternatively, it can be expressed solely in terms of the past values of Y_t:

$$Y_t \approx \sum_{j=1}^{d} c_j \cdot Y_{t-j} \tag{3}$$

To determine if X Granger-causes Y, we compare the forecast accuracy of the two models. If the first model provides a significantly better forecast than the second, we say that X Granger-causes Y.

To explore causality in the context of neural networks, we consider the multivariate time series $\mathbf{x}_t = (x_{t1}, ..., x_{tp})^T$. This series is used to predict another time series y_t by incorporating a set of neural network functions $\{g_1, ..., g_p\}$, where each function takes a segment of the past values of the vector $\mathbf{x}_t$. The predictive model can be expressed as:

$$y_t = g_1(x_{t-1}, ..., x_{t-K}) + ... + g_p(x_{t-1}, ..., x_{t-K}) + \varepsilon_t \tag{4}$$

Here, $g_i(\cdot)$ represents the neural network function approximating the relationship between the past values of the time series and the current value of y_t, and ε_t is the error term.

To assess causality within this framework, we compare the predictive power of a model including the neural network functions $g_i(\cdot)$ with a baseline model that excludes these functions. If the inclusion of $g_i(\cdot)$ significantly improves the prediction of y_t, we can infer a Granger-causal relationship from $\mathbf{x}_t$ to y_t.

Furthermore, if the predictive accuracy of the model is significantly enhanced by including neural network functions, it suggests the presence of a nonlinear relationship, which is a hallmark of Granger causality within the context of neural networks. This approach allows for a more nuanced understanding of causality, accounting for complex, nonlinear interactions that traditional methods may not capture.

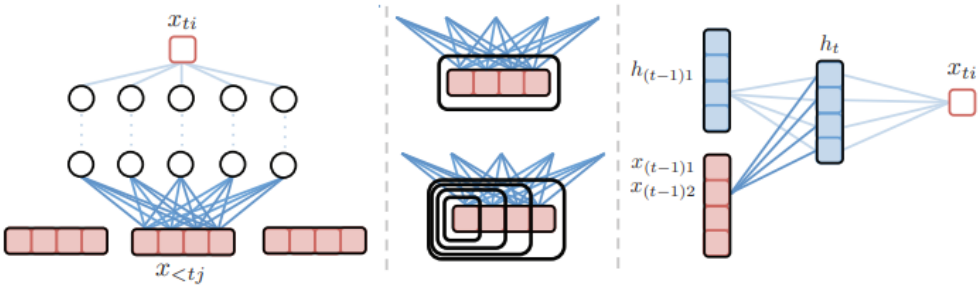


Figure 1: cMLP; group lasso penalty; cLSTM

3.4.1 cMLP

Consider an MLP with the function g_i, and the weight matrix $W = \{W^1, ..., W^L\}$, where each layer's weights W^l are applied to the inputs with a time lag K, along with a bias term to obtain the hidden layers h^l:

$$h_t^l = \begin{cases} \sigma\left(\sum_{k=1}^{K} W_k^l x_{t-k} + b^l\right), & l = 1 \\ \sigma\left(\sum_{k=1}^{K} W_k^l h_t^{l-1} + b^l\right), & l > 1 \end{cases} \tag{5}$$

To determine if the time series x_j Granger-causes the time series x_i, we can examine the significance of the weights in the MLP. If the weights W_{ij}^l connecting x_j to the prediction of x_i are non-zero, it suggests a potential Granger-causal relationship.

3.4.2 cLSTM

Consider an RNN with hidden state $\boldsymbol{h}_t \in \mathbb{R}^H$ update with some non linear function f_i as $\mathbf{h}_t = f_i(\mathbf{x}_t, \mathbf{h}_{t-1})$

To model complex temporal relationships effectively, we utilize an LSTM, which employs a cell state $\boldsymbol{c}_t$ alongside the hidden state $\boldsymbol{h}_t$ to update these states iteratively as

$$\begin{aligned} \mathbf{f}_t &= \sigma(W_f \mathbf{x}_t + U_f \mathbf{h}_{(t-1)}) \\ \mathbf{i}_t &= \sigma(W_i \mathbf{x}_t + U_i \mathbf{h}_{(t-1)}) \\ \mathbf{o}_t &= \sigma(W_o \mathbf{x}_t + U_o \mathbf{h}_{(t-1)}) \\ \mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \sigma(W_c \mathbf{x}_t + U_c \mathbf{h}_{t-1}) \\ \mathbf{h}_t &= \mathbf{o}_t \odot \sigma(\mathbf{c}_t). \end{aligned} \tag{6}$$

where $\odot$ denotes the element-wise multiplication and $\mathbf{i}_t$, $\mathbf{f}_t$ and $\mathbf{o}_t$ denote the input, forget and output gates and regulate the cell state's update, enabling long-range dependency capture for nonlinear time series analysis and prediction. The output for series i at time t is given as

$$x_{ti} = g_i(x_{<t}) + e_{ti} = W^2 h_t + e_{ti}. \tag{7}$$

Denote the full set of parameters as $\mathbf{W} = (W^1, W^2, U^1)$ and $W^1 = W^1 = \left((W^f)^\top, (W^i)^\top, (W^o)^\top, (W^c)^\top\right)$ and $U^1 = \left((U^f)^\top, (U^i)^\top, (U^o)^\top, (U^c)^\top\right)$. input matrices W^1 govern the impact of past observations on LSTM's gates and cell updates, thereby altering the hidden states.

3.4.3 Optimization Objective

Employing penalty promotes sparsity in $W^1$by:

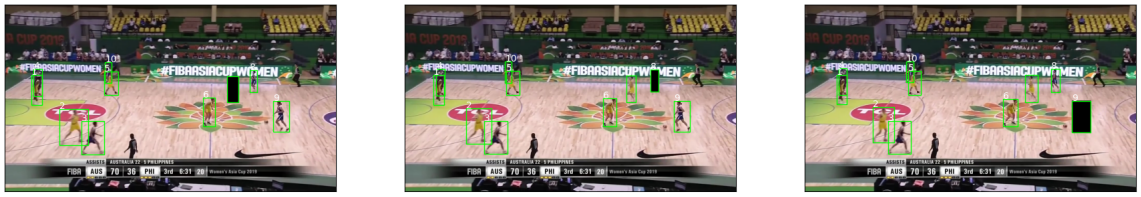
$$\min_{\mathbf{W}} \sum_{t=2}^{T} (x_{it} - g_i(x_{<t}))^2 + \lambda \sum_{j=1}^{p} \Omega(W_{:j}^1), \tag{8}$$

facilitating the selection of series with Granger-causal influence. For cMLP, the penalty could be a group lasso penalty $\|W_{:j}^1\|_F$, a group sparse group lasso $\Omega(W_{:j}^1) = \alpha \|W_{:j}^1\|_F + (1-\alpha) \sum_{k=1}^{K} \|W_{:j}^{1k}\|_2$ and $\Omega(W_{:j}^1) = \sum_{k=1}^{K} \left\|\left(W_{:j}^{1k}, \ldots, W_{:j}^{1K}\right)\right\|_F$. For cLSTM, the penalty could be group lasso penalty $\|W_{:j}^1\|_2$. A zero column in W^1 implies Granger non-causality for the corresponding input.

4 RESULTS

4.1 FASTER RCNN AND YOLOV8 DECTECTION RESULTS

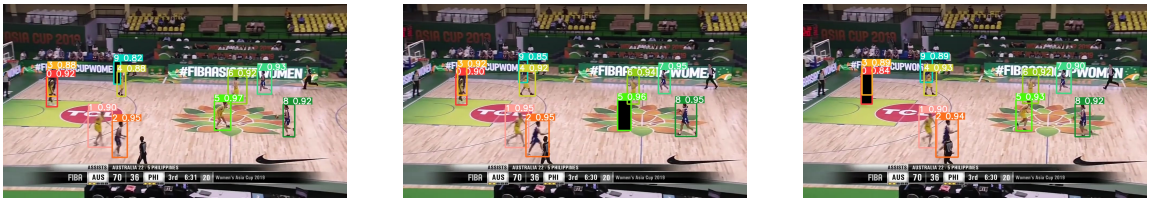
We employed the Faster R-CNN model on video frames with randomly masked objects to enable the model to learn interpolation, specifically predicting bounding boxes and labels:



(a) (b) (c)

Figure 2: The result of FasterRCNN for first few seconds with randomly masked input

Similar to FasterRCNN, YOLOv8 is also applied randomly masked, the results are shown as follows:



(a) (b) (c)

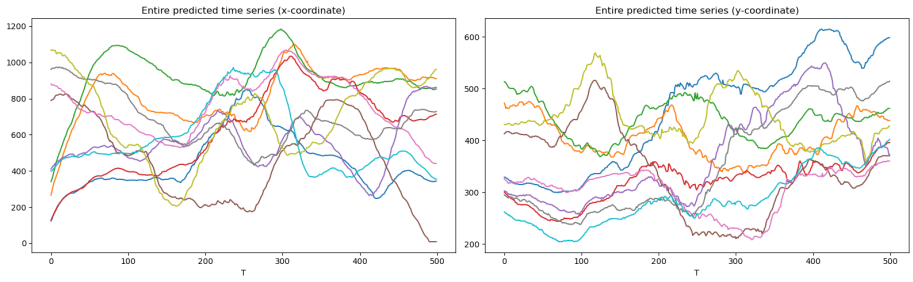
Figure 3: The results of YOLOv8 for a few seconds with randomly masked input

DeepSORT, utilizing YOLO for its detection backbone, showcases its tracking results as follows:

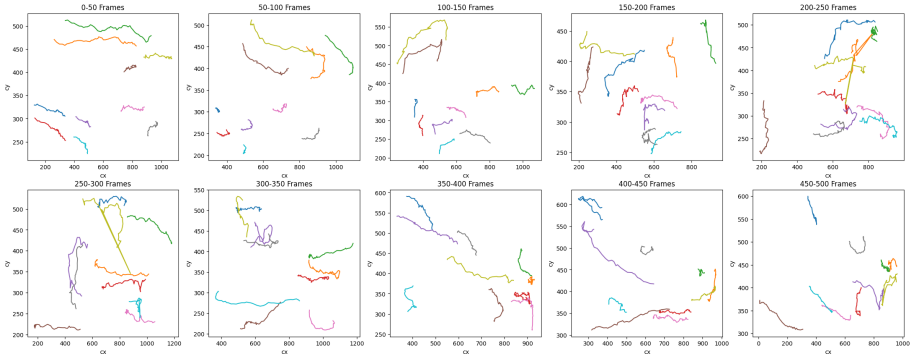
Figure 4: The result of DeepSORT for first few seconds

4.2 Neural Granger Causality Discovery

Post-object detection, we harnessed player central points to form time series data, detailing x, y, and center coordinates over time. The constructed series offers a visual insight into player movement dynamics.

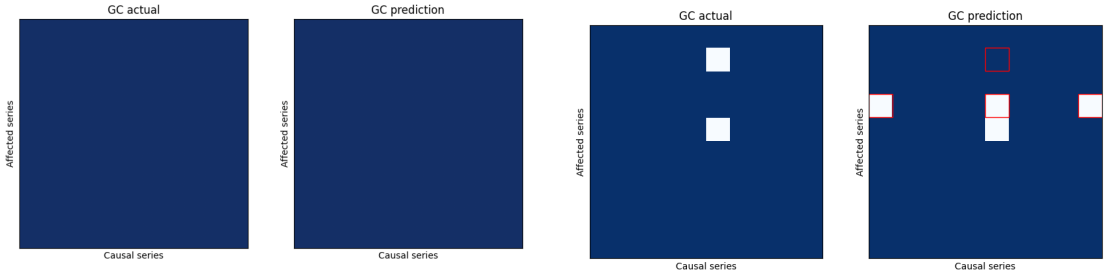


(a) x, y coordinates over time



(b) center coordinates over time

Utilizing cLSTM to unveil Granger causality relationships among players, we delved into the analysis by comparing true central point time series with predicted central point time series. The findings are highlighted in the following figure:



(a) cMLP results (b) cLSTM results

Figure 6: Granger Causality Matrix

Upon close examination, it becomes evident that the predicted causality, while not perfect, exhibits a high degree of accuracy, capturing a substantial portion of the causal relationships among players. It's important to note that while some causal connections may have been missed, the majority have been successfully identified.

5 Discussion and Conclusion

5.1 Mask

The paper by Van Dijk & De Croon (2019) reveals that MonoDepth primarily relies on the vertical position of objects for depth estimation. This preference suggests an underlying assumption of a flat ground and some knowledge of the camera's pose. In our work, we explore the effectiveness of masking techniques in depth estimation. Masking works effectively in this context because it helps isolate specific features or areas of an image, thus allowing the network to focus on relevant cues, such as vertical position, while ignoring irrelevant or misleading information. This approach enhances the network's ability to accurately interpret depth cues from a single image.

Furthermore, our experiments with both Faster-RCNN and YOLOv8 demonstrate their capabilities, albeit not perfect, in recognizing masked or blurred objects in videos post-training. This observation lends credence to the effectiveness of the masking approach, affirming its utility in enhancing object detection algorithms' depth perception accuracy.

5.2 Intersection Tracking

Intersection or occlusion remains a significant challenge in object tracking. The latest advancements, such as those outlined in Nasseri et al. (2021), build upon the DeepSORT algorithm, offering promising improvements. However, these methods still face limitations, particularly in dense and complex scenarios. The ongoing research in this field suggests that further enhancements in tracking algorithms, possibly integrating more sophisticated neural network architectures and better handling of object interactions, could offer substantial improvements in the future. This area remains ripe for exploration, with potential breakthroughs that could significantly advance the state of the art in object tracking.

5.3 Granger Causality

Harnessing the analytical might of cMLP and cLSTM models, we adeptly capture the nuanced causality woven into the fabric of a basketball game's dynamics. These models mirror the strategic interplay on the court, as players respond to the ball's movement, reflecting the game's inherent causality network. The near-universal detection of causality among players highlights their strategic maneuvers, whether in offense or defense, and paints a vivid picture of the game's complex interactions.

The variance observed between the cMLP and cLSTM results stems from several factors. The stochastic nature of Stochastic Gradient Descent (SGD) introduces variability in model outcomes. Additionally, each model's architecture leads to different interpretations of the non-linear temporal relationships between players, akin to varied tactical analyses of the same play.

Disparities may also arise from each model's unique approach to learning sequences and their sensitivity to hyperparameters, which affect causality pattern recognition. This not only emphasizes the sophistication of these learning techniques but also the layered complexity of causality in team sports, where each play is a tapestry of interdependent actions and reactions.

REFERENCES

Amy Fire and Song-Chun Zhu. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22, 2015.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, January 2023. URL `https://github.com/ultralytics/ultralytics`.

Mohammad Hossein Nasseri, Hadi Moradi, Reshad Hosseini, and Mohammadreza Babaee. Simple online and real-time tracking with occlusion handling. *arXiv preprint arXiv:2103.04147*, 2021.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

Tom Van Dijk and Guido De Croon. How do neural networks see depth in single images? In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2183–2191, 2019. doi: 10.1109/ICCV.2019.00227.

Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 748–756. IEEE, 2018. doi: 10.1109/WACV.2018.00087.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3645–3649, 2017. doi: 10.1109/ICIP.2017.8296962.