Enabling Visual Intelligence by Leveraging Visual Object States in a Neurosymbolic Framework

Filippos Gouidis [a,*], **Konstantinos Papoutsakis**[b], **Theodore Patkos**[a], **Antonis Argyros**[a] **and Dimtris Plexousakis**[a]

[a]Institute of Computer Science,
Foundation For Research and Technology,
Heraklion, Greece
[b]Department of Management,
Science Technology Hellenic Mediterranean University
Agios Nikolaos, Greece

Abstract. This paper investigates the potential of integrating visual object states for developing methods addressing complex visual intelligence tasks such as Long-Term Action anticipation (LTAA) and proposes that this should be achieved with the aid of a Neurosymbolic (NeSy) framework. We consider that this approach could offer significant advancements in applications requiring nuanced understanding and anticipation of future scenarios and could serve as an inspiration for the further development of Nesy methods exhibiting Visual Intelligence.

1 Introduction

Neurosymbolic artificial intelligence (NeSy AI) has developed significantly over the years, establishing itself as a major subfield of AI [5]. Traditionally, AI's neural and symbolic methods were considered to be in competition [4]. However, a recent surge in frameworks that combine these methods has been observed [9], driven by critical advancements of the limitations in deep learning [18, 15]. This interest continues even with the advancements made through scaling up deep learning, such as with large language models (LLMs), as researchers are highly motivated by the potential to leverage the strengths of both neural learning and symbolic knowledge representations. This synergy is increasingly regarded as a viable line of research toward achieving Artificial General Intelligence.

One of the primary challenges facing contemporary AI systems is their ability to handle complex tasks that require Visual Intelligence [6][19].

Notably, despite the fact that knowledge related to object states provides valuable cues toward this direction, the incorporation of this type of information within a NeSy framework has remained notably underexplored. We consider that object states hold significant potential to enhance the understanding and interaction with dynamic environments and can provide crucial context that can deepen an AI system's semantic comprehension and predictive capabilities (See Figure 1 for a characteristic example) and the integration of object states within a NeSy framework could bridge the gap between abstract symbolic reasoning and the fluidity of real-world neural perception, offering a more granular and accurate representation of real-time data.

The central thesis of our work [1] is that integrating object states within a NeSy framework is not only achievable but also greatly advantageous for improving the robustness and interpretability of various models tackling critical visual-based tasks. To illustrate this, we focus on Long Term Action Anticipation (LTAA), which constitutes a typical example of a problem where object state identification provides crucial information. Currently, the few LTAA methods that leverage knowledge about object states are exclusively data-driven and therefore we consider that the potential of using object states is under-explored. To support the above statement, we discuss the shortcomings of current methodologies and elaborate on the advantages that a NeSy approach exploiting knowledge related to object states can offer. Finally, we delineate the basic features of a novel NeSy framework that leverages object states.

2 The Problem

LTAA involves the prediction of future actions over long time-horizons in the range of several minutes from initial visual sequences in videos. This task is essential for applications where understanding and reacting to potential future scenarios are critical, such as autonomous driving, robotic assistants, and proactive healthcare systems. Unlike short-term prediction, which focuses on immediate next actions, long-term anticipation looks further ahead, often several minutes into the future. More formally: given a video sequence V consisting of frames $\{v_1, v_2, \ldots, v_T\}$, where T is the total number of frames, the objective is to predict a sequence of future actions A that are likely to occur following the last observed frame v_T. These actions are represented as $\{a_{T+1}, a_{T+2}, \ldots, a_{T+n}\}$, with n indicating the number of future steps to be anticipated.

* Corresponding Author. Email: gouidis@ics.forth.gr

[1] Acknowledgements: The Hellenic Foundation for Research and Innovation (H.F.R.I.) funded this research project under the 3rd Call for H.F.R.I. Research Projects to support Post-Doctoral Researchers (Project Number 7678 InterLinK: Visual Recognition and Anticipation of Human-Object Interactions using Deep Learning, Knowledge Graphs and Reasoning).

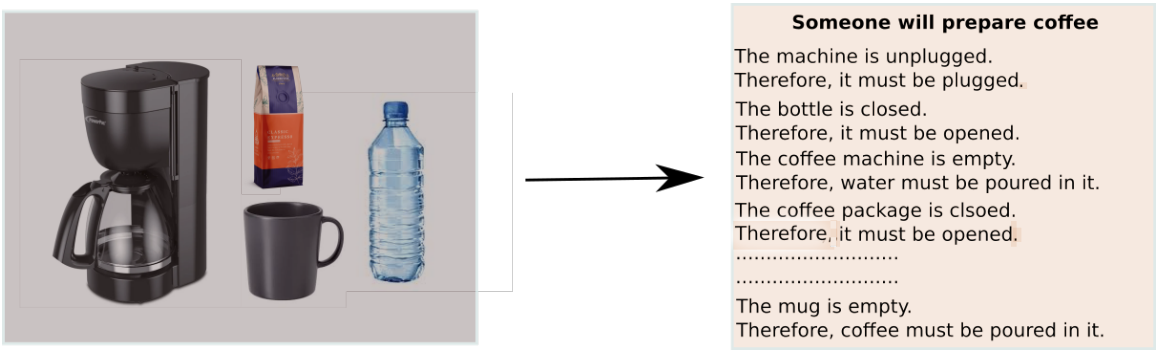

Figure 1. Knowledge related to object states is crucial for complex visual tasks such as LTAA. Consider a scene consisting of a coffee machine, a bottle of water, a package of coffee and a mug. The knowledge of the object classes allows the prediction of the activity, e.g. coffee preparation, but the knowledge of the object states is what enables the anticipation of the exact actions that constitute the activity.

3 Challenges

The task of LTAA involves several significant challenges, the most important of which are the following. Future relevant actions may be sparse and separated by long intervals of irrelevant activities or inactivity, making it difficult to pinpoint when significant actions will occur. Moreover, as the anticipation window increases, the dependency on a broader and more complex temporal context intensifies. Understanding which past cues are relevant for predictions far into the future requires sophisticated models that can integrate and reason over long temporal spans. Likewise, the farther into the future the prediction reaches, the higher the uncertainty. Multiple plausible futures can stem from a single point, and determining the most likely future action sequence can be highly ambiguous without extensive contextual understanding. Finally, LTAA often requires the integration of multiple data types (e.g., video, audio, textual descriptions) to accurately predict future actions, adding to the complexity of model design and training.

4 Limitations of current approaches

Very few of the works addressing the LTAA problem are based on the NeSY paradigm [2, 1, 11]. However, none of these methods utilizes information pertaining to object states in any way. The vast majority of the LTAA methods follows a data-driven approach (some notable works include [14, 16, 7, 12, 3, 8, 13, 10, 17]). Although some of these works use object-states related information this is being done in a purely data-driven manner. Moreover, these methods suffer from a number of important shortcomings.

Namely, due to their reliance on purely data-driven approaches these methods struggle with the high dimensionality and variability inherent in video data. Moreover, data-driven approaches rely heavily on the availability of large, annotated datasets that adequately represent the diversity of possible scenarios. However, such datasets are often scarce or biased. Furthermore, these methods typically require substantial computational resources for training and inference. Additionally, these purely statistical models often fail to capture the causal relationships and complex interactions between objects and humans in dynamic environments, They tend to predict future actions based on correlations observed in the training data, without a genuine understanding of the underlying causal mechanisms, a limitation becomes particularly evident in scenarios where contextual understanding and reasoning about object states are crucial.

5 Leveraging Object States within a NeSy Framework

Integrating symbolic reasoning with object states allows models to apply logical rules and knowledge of object interactions to predict future actions. This approach offers a rich contextual understanding absent in purely data-driven systems, facilitating the interpretation of complex scenarios and enabling predictions based on causal relationships rather than mere correlations. Symbolic components enhance the system's ability to reason about actions with minimal data by leveraging predefined rules and extensive knowledge bases. This capability is particularly beneficial in situations with limited or biased training data, enabling the model to generalize from fundamental object interaction principles. Object states provide clear environmental information, reducing uncertainty in long-term predictions. Neuro-symbolic frameworks incorporating object states offer significant adaptability, allowing for the integration of new rules and knowledge to accommodate novel or changing environments. This adaptability is essential for applications requiring operation under diverse conditions. Moreover, the use of object states enhances the explainability of model predictions. Actions predicted based on clear, traceable rules related to changes in object states are more understandable and verifiable by users, fostering trust in the model's outputs. Additionally, object state information can direct the focus of video processing, prioritizing sections with significant state changes. This selective attention can reduce computational demands and increase the efficiency of the model. Finally, leveraging object state changes as additional features enables models to learn effectively from fewer examples minimizing the need for large labeled datasets and reducing the risk of bias inherent in the training data.

6 Proposed Approach

We present the basic outline of a Nesy framework that leverages visual object states analysis. Overall, the proposed model comprises three main components: a neural module, a symbolic module and a reasoning engine. A sketch of the model along with the corresponding information flow is shown in Figure 2.

Neural Component for Perception: The neural component encompasses the following modules: an object detector (OD), an object tracker (OT), a Relation Detector (RD), a state classifier (SC) and a action recognizer (AR). The OD and OT are responsible for classifying and tracking the instances of object classes across the video respectively. The RD focuses on discerning the relationships between

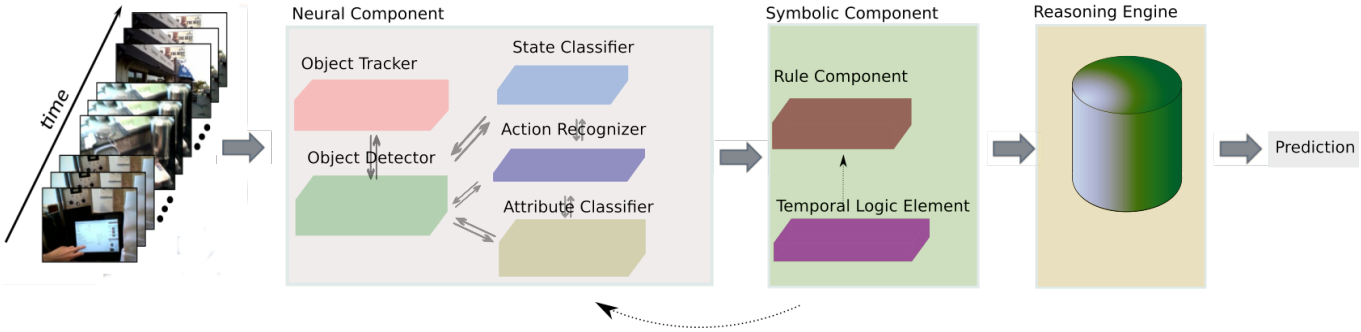

Figure 2. A schema of the framework we are proposing.

Module	Type of Input	Type of Output	Example Method/Network
Object Detector	RGB Images	Labels and Bounding Boxes	YOLO, Faster R-CNN
Object Tracker	RGB Images	Object IDs	SORT
Relation Detector	RGB Images	Interaction/ Objects Relationships Labels	Graph Convolutional Networks
State Classifier	RGB Images	State Labels	ResNet, VGG
Action Recognizer	RGB Images	Action Labels	I3D (Inflated 3D ConvNet)
Rules Component	State Labels	State Transitions	Custom Rule-Based System
Temporal Logic Element	State Transitions	Temporal Constraints	Temporal Logic Networks
Reasoning Engine	All Outputs	Predicted Actions	Prolog, Answer Set Programming (ASP)

Table 1. Modules of the Pipeline in the Proposed Neuro-Symbolic Framework

different objects within the scene, capturing interactions that are critical for understanding the context. The SC is used to infer the current states of the observed objects, which is vital for interpreting how these states might influence the subsequent actions. Finally, the AR module is responsible for recognizing the actions.

Symbolic Component for Structured Representation: The symbolic component of the model consists of the following modules: a Rules Component (RC) and a Temporal Logic Element (TLE). The RC encompasses a set of rules that govern state transitions, object affordances, and the pre-conditions and effects associated with various actions. It systematically codifies how changes in object states and interactions lead to different outcomes within the video context. The TLE is equipped with specialized knowledge pertaining to the temporal alignment and constraints related to the objects, their states, and the resultant actions. This module ensures that the timing and sequence of events are accurately maintained and logically consistent throughout the analysis.

Reasoning Engine: The reasoning engine functions within a logic programming framework, designed to predict long-term actions by leveraging the outputs from the symbolic component. This framework systematically processes and interprets the symbolic representations of object interactions and state transitions, enabling it to anticipate future activities by applying established logical rules and relationships. Through this integration, the engine effectively synthesizes the insights gained from the symbolic component to forecast actions that may occur in extended future scenarios.

Pipeline: The neural component initially processes the video sequences, with each of its specialized modules dedicated to specific perceptual tasks. The outputs generated by these modules are then forwarded to the symbolic component of the system. Based on these inputs, specific rules within the RC are triggered depending on the contextual data and object interactions identified. These activated rules are subsequently integrated with the temporal information encapsulated in the TLE. The reasoning engine serves as the decision-making core of the model, utilizing both the static and dynamic aspects of the inputs to generate predictions about future actions.

Functional Specifications of Pipeline Modules: In the following we present a rudimentary specification concerning the inputs and outputs of the different modules along with examples of available methods and neural networks that can support the necessary functionality (a summary of the specification is shown in Table: 1.). The OD module is to take RGB images as input and outputs labels and bounding boxes to identify and localize objects within each frame. For this module standard off-the-shelf networks such as YOLO or Faster R-CNN could be utilized. The OT would take as input RGB images and output unique IDs to objects across frames. A suitable method that could be employed in this context is the SORT tracking algorithm.

The RD module is to take as an input RGB images and produce labels corresponding to interactions and relationships between objects within the scene. A straight-forward way to achieve this is through the use of Convolutional Networks (GCNs). The SC would take RGB images as input and output state labels. Again the utilization of GCNslike ResNet or VGG seems as the most apropriate choise.

The AR is to processes RGB images and generate action labels. A network that could support is the Inflated 3D ConvNet (I3D). The RC would be given as input state labels and generated predictions for state transitions. This could be achieved by employing custom rule-based systems to codify the logical rules governing state changes and object interactions. The TLE is to take the previous state transitions and output temporal constraints. For this task utilizing Temporal Logic Networks seems as the most suitable option. Finally, the RE would synthesize all the previous outputs, including labels, state transitions, and temporal constraints, in order to produce predictions for future actions. Logic programming frameworks such as Prolog or Answer Set Programming (ASP) support this functionality and appear as the most appropriate options.

7 Conclusion

This paper investigates the potential of integrating visual object states into a NeSy framework in the context of complex Visual Intelligence Tasks such as LTAA. By bridging the gap between symbolic reasoning and neural perception of real-world dynamics, the proposed NeSy framework aims to provide a more granular and accurate representation of environments and considerably improve the predictive capabilities. We hope that this study will serve as an initial step towards uncovering the potential offered by this approach.

Acknowledgements: The Hellenic Foundation for Research and Innovation (H.F.R.I.) funded this research project under the 3rd Call for H.F.R.I. Research Projects to support Post-Doctoral Researchers (Project Number 7678 InterLinK: Visual Recognition and Anticipation of Human-Object Interactions using Deep Learning, Knowledge Graphs and Reasoning).

References

[1] N. Bellotto, L. Castri, M. Hanheide, and S. Mghames. A neuro-symbolic approach for enhanced human motion prediction. *repository.lincoln.ac.uk*, 2023.

[2] S. Bhagat, S. Stepputtis, and J. Campbell. Knowledge-guided short-context action anticipation in human-centric videos. *arXiv preprint arXiv:2309.05943*, 2023.

[3] S. Das and M. Ryoo. Video+ clip baseline for ego4d long-term action anticipation. *arXiv preprint arXiv:2207.00579*, 2022.

[4] L. De Raedt, S. Dumančić, R. Manhaeve, and G. Marra. From statistical relational to neuro-symbolic artificial intelligence. *arXiv preprint arXiv:2003.08316*, 2020.

[5] A. d. Garcez and L. C. Lamb. Neurosymbolic ai: The 3 rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, 2023.

[6] D. Geman, S. Geman, N. Hallonquist, et al. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623, 2015. doi: 10.1073/pnas.1422953112. URL https://www.pnas.org/doi/abs/10.1073/pnas.1422953112.

[7] D. Gong, J. Lee, M. Kim, and S. Ha. Future transformer for long-term action anticipation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4377–4386, 2022.

[8] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[9] P. Hitzler, A. Eberhart, and M. Ebrahimi. Neuro-symbolic approaches in artificial intelligence. *National Science Review*, 9(6), 2022.

[10] D. Huang, O. Hilliges, L. Van Gool, and X. Wang. Palm: Predicting actions through language models@ ego4d long-term action anticipation challenge 2023. *arXiv preprint arXiv:2306.16545*, 2023.

[11] M. Katz, K. Srinivas, and S. Sohrabi. Scenario planning in the wild: A neuro-symbolic approach. In *Proceedings of the FinPlan Workshop at ICAPS*, 2021.

[12] Q. Ke, M. Fritz, and B. Schiele. Time-conditioned action anticipation in one shot. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11004–11013, 2019.

[13] E. Mascaró, H. Ahn, and D. Lee. Intention-conditioned long-term human egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2164–2173, 2023.

[14] M. Nawhal, A. Jyothi, and G. Mori. Rethinking learning approaches for long-term action anticipation. In *European Conference on Computer Vision*, pages 123–139. Springer, 2022.

[15] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387. IEEE, 2016. URL https://ieeexplore.ieee.org/document/7467366.

[16] C. Patsch, J. Zhang, Y. Wu, and M. Zakour. Long-term action anticipation based on contextual alignment. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[17] S. Thakur, C. Beyan, and P. Morerio. Leveraging next-active objects for context-aware anticipation in egocentric videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[18] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, 2020. URL https://arxiv.org/abs/2007.05558.

[19] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.