A Hybrid Approach for Visual Recognition of Object States

Filippos Gouidis

Computer Science Department, University of Crete, Heraklion, Greece
Institute of Computer Scicence, Foundation for Research and Technology, Heraklion, Greece
gouidis@csd.uoc.gr

Abstract

The basic objective of my research work is to address the challenging problem of recognizing object states in a visual context by integrating data-driven and symbolic approaches. In particular, I focus on the Zero-shot variation of this task. Key contributions include the development of novel methods that exhibit state-of-the-art (SOTA) performance, the creation of a new object states dataset, the formulation of novel problems, the successful integration of low-level and high-level approaches, and comprehensive analyses that highlight the specific challenges posed by the problem.

Research Problem

The main topic of my dissertation is the visual recognition of object states. In this context, I aim to develop methods that are able to classify the states of the objects appearing in images and/or videos. In particular, I focus on the zero-shot variation of this problem, in which no visual information related to object states is available for training. The core objective of my research is to explore efficient approaches for the integration of data-driven (low-level) and symbolic (high-level) approaches.

The problem of Object State Classification (OSC) is crucial in our daily interactions with various objects in different contexts. Recognizing the states of objects is essential as it determines an object's condition and the actions that can be performed with it (Jamone et al. 2016). In computer vision, OSC is closely related to action recognition (Wang, Farhadi, and Gupta 2016), object detection and classification (Farhadi et al. 2009), and affordance learning (Chuang et al. 2018). Moreover, states also provide cues on the dynamic aspects and transformations of objects, which are crucial for action recognition in images and videos (Liu, Wei, and Zhu 2017).Despite its importance, research on OSC has been limited compared to object classification. However, recent years have seen an increase in studies dedicated to this problem (Souček et al. 2022; Gouidis et al. 2022).

In addition to the standard challenges inherent in every computer vision problem, the task of OSC poses some unique challenges. First, states are more difficult to recognize than object classes or attributes because they involve a more complex representation of visual information. Classes and attributes are usually defined based on visual properties that remain relatively stable across different contexts and appearances, such as color, texture, shape, or size. In contrast, states are defined based on changes in appearance or context, which are more subtle and can be influenced by many factors.Additionally, states are often more context-dependent and task-specific, meaning they may not be applicable or meaningful in all contexts. Therefore, recognizing states involves more complex reasoning and inference than recognizing attributes, requiring models that can capture and integrate both visual and semantic information from the scene.

Contributions

The main contributions of my research are the following.

1. Extensive Analysis and Dataset Creation: An in-depth analysis of OSC was conducted, investigating several aspects of the problem and studying it in conjunction with the closely related problem of object recognition (Gouidis et al. 2022). This comprehensive analysis led to several key findings, the most significant being that OSC cannot be addressed with the standard methods employed for object classification. An important outcome of this study was the creation of a new publicly available Object States Dataset[1].

2. Introduction of Zero-Shot Object-Agnostic State Classification: A novel variation of the OSC problem, i.e., the Zero Shot Object-Agnostic State Classification (ZS-OASC) task was formulated. This task deals with the recognition of object states in images and/or videos when no visual training samples (zero shot) and cues related to object classes (object-agnostic) are utilized. This approach differs from the standard strategy for addressing OSC, which typically relies on accurate object classification. This significantly broadens the applicability of OSC by eliminating the reliance on object-specific training data.

3. Hybrid Method for ZS-OASC: We developed a hybrid method specifically designed for ZS-OASC (Gouidis et al. 2023). This method combines neural modules, i.e., Convolutional Neural Networks (CNNs) and Graph Convolutional Networks (GNNs), with symbolic components, such

[1] https://github.com/philipposg/OSDD

as Knowledge Graphs (KGs) in the context of projecting semantic knowledge into the visual space. This enables the generation of visual embeddings for any state class, making the method suitable for zero-shot classification. Our approach achieved state-of-the-art performance across all available object states datasets.

4. Exploration of the potential of KGs for OSC: A novel method for the effective and efficient utilization of KGs as a knowledge source for zero-shot visual tasks was introduced (Gouidis et al. 2024c). This study led to the development of a technique that enables the semi-automatic construction of domain-specific KGs, drawing knowledge from various semantic sources. An important finding was the impact of different types of knowledge regarding the suitability of generated embeddings for the ZS-OASC task. Another key discovery was the adverse effect of noise in domain-agnostic KGs used for training GNNs on the overall performance of the method.

5. Large Language Models Integration: More recently, a significant breakthrough involved the utilization of Large Language Models (LLMs) for generating semantic features (Gouidis et al. 2024a). This study involved an extensive investigation into combining domain-specific and general-purpose knowledge in the context of which special methods were devised to generate a domain-specific corpus through tailored LLMs queries and combine it with general-purpose corpora available on the Web. This approach improved significantly the performance of our hybrid method for the ZS-OASC task. Furthermore, the use of LLMs for the construction and refinement of KGs led to further performance gains (Gouidis et al. 2024b).

6. State Change Anticipation in Videos: We introduced the novel problem of State Change Anticipation in Videos (SCAV) (Manousaki et al. 2024) which focuses on predicting state changes of objects in videos. This task has significant implications for scene understanding, automated monitoring, and action planning. It integrates learned visual features with natural language processing (NLP) features representing past object state changes. Extensive experimental evaluation demonstrates the method's effectiveness in predicting object state changes in dynamic scenarios, highlighting the potential of combining video and linguistic cues to enhance predictive performance.

Future Directions

In the remaining time of my PhD studies, I plan to conclude my research by building on my previous work. Specifically, I aim to explore the potential of leveraging meta-paths to enhance embedding generation. Current approaches treat KGs as homogeneous sources, overlooking the rich, heterogeneous information embedded in their diverse nodes and relations. By utilizing meta-paths, I intend to develop methods for more efficient processing of this complex, structured data. Moreover, I would like to experiment is the utilization of multi-modal KGs in the context of ZS-OASC and to extend and refine the existing work concerning the problem of State Change anticipation in Videos.

References

Chuang, C.-Y.; Li, J.; Torralba, A.; and Fidler, S. 2018. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 975–983.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, 1778–1785.

Gouidis, F.; Papantoniou, K.; Papoutsakis, K.; Patkos, T.; Argyros, A.; and Plexousakis, D. 2024a. Fusing Domain-Specific Content from Large Language Models into Knowledge Graphs for Enhanced Zero Shot Object State Classification. In *Proceedings of the AAAI Symposium Series*, volume 3, 115–124.

Gouidis, F.; Papantoniou, K.; Papoutsakis, K.; Patkos, T.; Argyros, A.; and Plexousakis, D. 2024b. LLM-aided Knowledge Graph construction for Zero-Shot Visual Object State Classification. 611–619.

Gouidis, F.; Papoutsakis, K.; Patkos, T.; Argyros, A.; and Plexousakis, D. 2024c. Exploring the Impact of Knowledge Graphs on Zero-Shot Visual Object State Classification. In *Proceedings of the the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications.*

Gouidis, F.; Patkos, T.; Argyros, A.; and Plexousakis, D. 2022. Detecting Object States vs Detecting Objects: A New Dataset and a Quantitative Experimental Study. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, volume 5, 590–600.

Gouidis, F.; Patkos, T.; Argyros, A.; and Plexousakis, D. 2023. Leveraging knowledge graphs for zero-shot object-agnostic state classification. *arXiv preprint arXiv:2307.12179*.

Jamone, L.; Ugur, E.; Cangelosi, A.; Fadiga, L.; Bernardino, A.; Piater, J.; and Santos-Victor, J. 2016. Affordances in psychology, neuroscience, and robotics: A survey. *IEEE Transactions on Cognitive and Developmental Systems*, 10(1): 4–25.

Liu, Y.; Wei, P.; and Zhu, S.-C. 2017. Jointly recognizing object fluents and tasks in egocentric videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2924–2932.

Manousaki, V.; Bacharidis, K.; Gouidis, F.; Papoutsakis, K.; Plexousakis, D.; and Argyros, A. 2024. Anticipating Object State Changes. *arXiv preprint arXiv:2405.12789*.

Souček, T.; Alayrac, J.-B.; Miech, A.; Laptev, I.; and Sivic, J. 2022. Multi-Task Learning of Object State Changes from Uncurated Videos.

Wang, X.; Farhadi, A.; and Gupta, A. 2016. Actions $\sim$ Transformations. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, 2658–2667. IEEE. ISBN 9781467388504.