Efficient Object Detection Through Proto-object Selection

Caterina Caccavella[1,*], Vittorio Fra[1,2], Andreas Ziegler[3], Giulia D'Angelo[4], Yulia Sandamirskaya[1]

[1] Institute of Computational Life Sciences, Zurich University of Applied Sciences (ZHAW), Wädenswil, Switzerland
[2] Politecnico di Torino, Turin, Italy
[3] University of Tübingen, Tübingen, Germany
[4] Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University, Prague, Czech Republic

Abstract Conventional frame-based cameras suffer from high data redundancy and limited temporal resolution, making them inefficient for real-time tasks in dynamic environments. To address these limitations, this work proposes a real-time, event-based object detection framework grounded in the fundamental assumption that objects are continuous and close entities in space. The use of event-based cameras, inspired by the human retina, minimize latency, energy consumption, and data redundancy while supporting high dynamic range perception. Moreover, event input naturally extracts edges in the scene, a crucial feature for object identification. A biologically inspired selective attention mechanism further reduces data processing by dynamically selecting regions of interest (ROIs) in the sparse input signal that may contain objects. The proposed framework uses a modular architecture that includes a saliency-based attention model, a lightweight classifier, and two Dynamic Neural Fields (DNFs), used respectively for selecting the ROI in the scene and for implementing a scene memory module. The first DNF integrates input from the attention model, previously attended features, and input events to select the ROI through dynamic competition among multiple salience peaks. The lightweight classifier, designed with a minimal number of parameters for fast training and deployment, classifies the content within the ROI. The output is stored in a second DNF, which maintains a memory of recognized objects and their locations. A real-time demonstration illustrates the system's ability to recognize objects in an open-world scenario, emphasizing the benefits of combining learning-free, low-latency, and low-power proto-object extraction and lightweight classifiers.

Methods An event-based camera, mounted on a pan-tilt robotic unit that serves as the robot's head, observes a tabletop scene containing various objects. The event stream from the camera is processed by the proto-object model to detect salient regions in the scene [1]. The pixel coordinates and approximate size of the most salient region define the spatial region of interest, which is fed as input to the DNF, together with the input from the camera. The DNF receives excitatory input from the sensor's event stream and the saliency map, as well as inhibitory input from the location of the previous ROI. By evolving toward a stable peak of activation, the DNF selects the ROI, whose content is then passed to a lightweight object classifier trained on isolated, centered objects with minimal background. The resulting object locations are stored in the working memory module, implemented using the DNF framework and inspired by [2]. The locations of recognized object representations, together with the saliency mechanism, are used to generate corresponding gaze shifts for the robot, facilitating object recognition and scene understanding. An overview of the system architecture is shown in fig. 1.

Preliminary Results Initial results include integrating the attention model with the event-based camera and preparing preliminary test datasets and classifiers. The attention model was evaluated on the synthetic circle dataset IROSsalmap [3] to assess its response to varying object sizes, helping define the operative range of scales. The evaluation metric, based on detection accuracy, is defined as the number of correctly identified circles that are qualitatively selected, divided by the total number of objects present in the scene. Properly tuned parameters allowed the selection of 5 out of 6 circles in the characterization test. To test the pipeline on more realistic data, we selected 4 object classes from the YCB-ev dataset [4] and recorded a short data sequence. Both datasets were preprocessed into two variants: one with cropped object regions (to simulate the ROI selected by the attention model), and another with the full scene preserved to allow comparison with standard methods. Initial tests trained a 2-layers MLP on four YCB-ev objects, achieving 96.15% accuracy. Current work includes applying the attention model to selected datasets and initiating the first coupling with the lightweight object classifier, together with the implementation of the DNF modules.

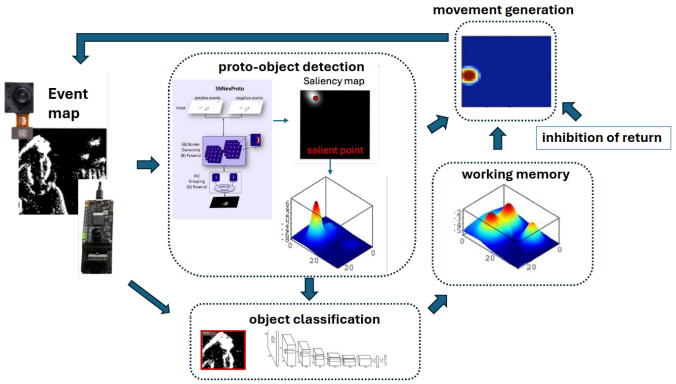

Figure 1: System architecture overview: Event-based camera input is fed to the attention model and to the first DNF. The DNF receives as input the saliency-map, the event input and the location of already recognized objects, and selects the area of the scene to pass to the lightweight object classifier. The recognized object is stored in the working memory (second DNF), which is used to generate gaze shifts and modulate the first DNF.

References

[1] G. D'Angelo, A. Perrett, M. Iacono, S. Furber, and C. Bartolozzi, "Event driven bio-inspired attentive system for the icub humanoid robot on spinnaker," *Neuromorphic Computing and Engineering*, vol. 2, no. 2, p. 024008, 2022.

[2] R. Grieben, S. Sehring, J. Tekülve, J. P. Spencer, and G. Schöner, "Roboverine: A human-inspired neural robotic process model of active visual search and scene grammar in naturalistic environments," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 11 470–11 477.

[3] M. Iacono, G. D'Angelo, A. Glover, V. Tikhanoff, E. Niebur, and C. Bartolozzi, "Proto-object based saliency for event-driven cameras," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 805–812.

[4] P. Rojtberg and T. Pöllabauer, "Ycb-ev 1.1: Event-vision dataset for 6dof object pose estimation," in *European Conference on Computer Vision*. Springer, 2025, pp. 1–13.