Object Detection for Embodied Agents using Sensory Commutativity of Action Sequences

Hugo Caselles-Dupré, Michael Garcia Ortiz, David Filliat

Object Detection for Embodied Agents using Sensory Commutativity of Action Sequences

Hugo Caselles-Dupré[1,2], Michael Garcia-Ortiz[3], David Filliat[1]

[1]Flowers Laboratory (ENSTA Paris & INRIA)
[2]AI Lab (Softbank Robotics Europe)
[3]City of London University
casellesdupre.hugo@gmail.com, michael.garcia-ortiz@city.ac.uk, david.filliat@ensta.fr

Abstract

We study perception in the scenario of an embodied agent equipped with first-person sensors and a continuous motor space with multiple degrees of freedom. We introduce a method for movable and immovable object detection that exploits the commutative properties of action sequences. Our method is based on playing an action sequence in two different orders from the same starting point, and tested on 3D realistic robotics setups.

1 Introduction

Perception is the medium by which agents organize and interpret sensory stimuli, in order to reason and act in an environment using their available actions [7]. We focus on scenarios where embodied agents are situated in *realistic* environments, i.e. the agents face partial observability, coherent physics, first-person view with high-dimensional state space, and low-level continuous motor (i.e. action) space with multiple degrees of freedom. These embodied agents, when acting in such an environment, produce a stream of sensorimotor data, composed of successions of motor states and sensory information. While most current approaches for building perception focus on studying the sensory information alone, several approaches [2, 11, 5, 15, 1] that can be traced back to 1895 [14], advocate the necessity of studying the relationship between sensors and motors for the emergence of perception.

The present work introduces an object detection method based on previous work on sensory commutativity of action sequences for embodied agents [2]. Starting from the sensory commutativity experiment (SC-experiment), which consists in playing an action sequence in two different orders from the same starting point, we aim at extracting information about what the agent can interact with in the environment. We thus introduce a method that shows it is possible to learn about immovable and movable objects in 3D realistic environments, and with promising generalization properties.

2 Related work and motivation

Sensory commutativity of action sequences. Our work is based on the study of commutativity of action sequences with respect to sensors, termed sensory commutativity. The Sensory Commutativity-experiment (SC-experiment), which consists in having the agent play an action sequence in two different orders from the same starting point and comparing the two outcomes in observations (are the observations equal or not?), is the basis for studying sensory commutativity. From this experiment, the Sensory Commutativity Probability (SCP) has been defined as the probability that a sequence

NeurIPS 2020 Workshop on BabyMind, Vancouver, Canada.

of movements using only one degree of freedom of the agent, an arm joint, for instance, sensory commutes. It has been showed that this value has meaning for the embodied agent: if the SCP is high then the degree-of-freedom has a low impact on the environment (e.g. moving a shoulder is more likely to lead to environment changes than moving a finger, so SCP for a shoulder is lower than for a finger).

Object detection for embodied agents. Object detection is a large body of work that is related to the present work. We do not claim to improve over state-of-the-art results in this field. Our main motivation is rather to give insights on how sensory commutativity can allow seeing the problem in a novel way. For object detection, we either have well-performing methods based on computer vision algorithms and largely annotated databases [6], or algorithms based on data collected by the agent itself [3, 12, 9]. With sensory commutativity, we fall in the second category, as we aim at using sensory commutativity as the tool for detecting objects that the agent can interact with.

3 Sensory Commutativity Probability for object detection

Our experiments are performed in iGibson [17] with the Fetch robot. iGibson is a simulation environment for robotics providing fast visual rendering and physics simulation. It is packed with a dataset with hundreds of large 3D environments reconstructed from real homes and offices, and interactive objects that can be pushed and actuated. In our experiments, we use the Rs environment, which is basically a regular apartment. We place the Fetch robot in this environment. Fetch is originally a 10-DOF real robot [16] equipped with a 7-DOF articulated arm, a base with two wheels, and a liftable torso. Fetch perceives the environment through a camera placed in his head.

3.1 Intuition

The concept of SCP is based upon comparing outcomes of SC-experiments and evaluating whether the two resulting observations are considered equal or not. Going beyond this equality test, we propose to have a finer analysis of the differences between the two observations obs_1 and obs_2 resulting from an SC-experiment.

We argue that comparing obs_1 and obs_2 leads to three possible outcomes from which the agent can learn about immovable and movable objects in the environment.

- obs_1 and obs_2 are different: the two action sequences from this starting position do not commute, because the robot interacted with immovable objects. Using the position of the agent, we can now map immovable objects in the environment.
- obs_1 and obs_2 are identical: the two action sequences from this starting position commute, because the robot did not interact with anything in the environment (free movement). Using the position of the agent, we know that there are no objects in the current space around the robot.
- obs_1 and obs_2 are identical except for an object that has been moved: it's the case where the robot has interacted with a movable object that did not block the robot's movement. Hence the two action sequences did commute, except for the object that has been moved. We can learn to detect this moving object and track it.

3.2 Method

In order to verify the aforementioned intuition, Fetch needs to be able to perform an SC-experiment and then detect: 1) if the two resulting observations are identical or not, 2) if they are identical except for the parts of an image corresponding to an object that moved. For that, we equip the agent with a vision system that gets two observations as input and outputs two masks which will be all zeros if the two observations are identical, all ones if they are different, and the mask of an object if this object moves. Studies in cognitive science indicate that children are capable of doing this differentiation at a very young age (1 month old) [10, 8], so we believe equipping the agent with this basic ability is a reasonable assumption.

We thus train a neural network with generated data to predict those two masks with two observations as input. We refer to this model as the "mask predictor".

Dataset. The data is collected in the Placida environment by starting at a random position in the environment (observation obs_1) and then collecting data for the three possible outcomes:

- no difference: it suffices to keep the same observation and the corresponding masks are all zeros. The data is (obs_1 + all zeros mask, obs_1 + all zeros mask).
- completely different: we move the robot and get a different observation obs_2, the corresponding masks are all ones. The data is (obs_1 + all ones mask, obs_2 + all ones mask).
- no difference except moved objects: we randomly disturb the orientation and position of some movable objects and get a new observation obs_2 identical to obs_1 up the moved objects. The data is (obs_1 + moving objects mask, obs_2 + moving objects mask)

The resulting dataset is illustrated in Fig.1.

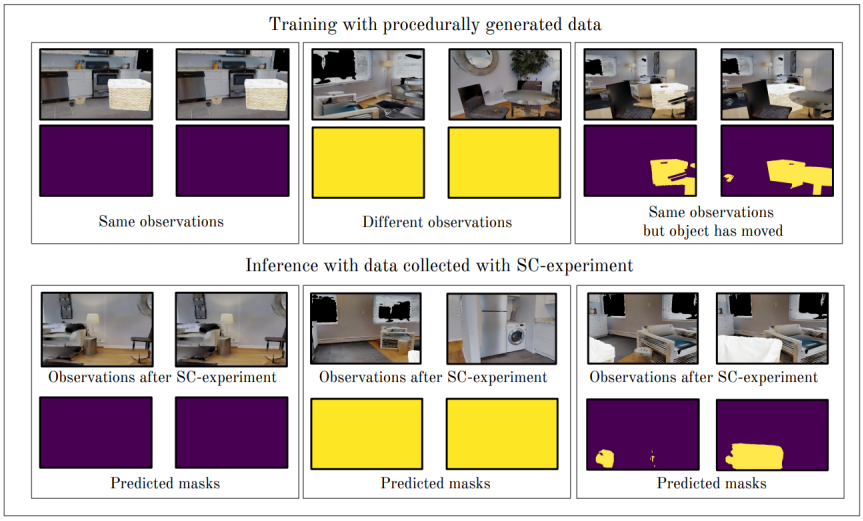

Figure 1: Dataset for training the mask predictor and inference results on data collected with SC-experiments. The dataset is procedurally generated to simulate the three possible scenarii resulting from a SC-experiment. **Left:** scenario where there are no changes in the observations. **Middle:** scenario where the observations are different. **Right:** scenario where the observations are identical up to moved objects.

Architecture and training. We then train the neural network to predict the masks given the observations. This process is similar to predicting the optical flow of two consecutive frames in a video. Thus, for the mask predictor, we use the architecture of FlowNet-S [4], a popular baseline for optical flow prediction. We train the model using the same architecture and optimization process, except we change the loss function to a binary cross-entropy loss between the ground truth mask and the output mask of the network. All training details are available in the open-source implementation we used [1].

Inference. Once the mask predictor is trained, we place the agent in an environment and perform SC-experiments where we let it play an action sequence in different orders from the same starting point. Then, the goal is for the agent to detect immovable and movable objects using the generated data from the SCP-type experiment and the mask predictor. All experimental details are described in App.A.

3.3 Experiments and results

Do SC-type experiments allow detecting and track movable objects? We compute SC-experiments using a DOF selected using SCP value. We select the DOF with SCP closest to 0.5 in order for the outcome of SC-experiments to be as diverse as possible, i.e. the DOF of the arm that is closest to the body of the agent. Results presented in Figs.1 & 2 show that using the

[1] `https://github.com/ClementPinard/FlowNetPytorch`

mask detector with the outcome of these SC-experiments allows to detect objects that have been moved. Note that the mask detector only detects objects that have moved between the two resulting observations, rightfully ignoring the other potential objects that were not moved. After this detection, we can then use semi-supervised tracking algorithms such as [13] in order to track the detected object.

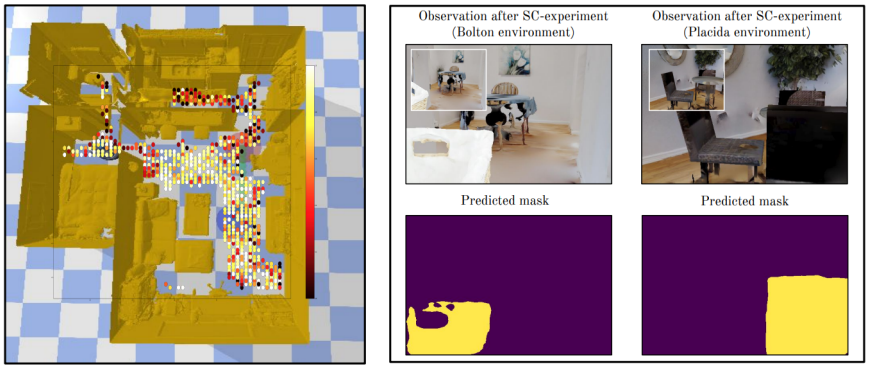

Figure 2: **Left:** Local SCP value corresponding to the arm's DOF that is closer to the body of the agent, computed over 10 SC-experiments for each position in the Rs-interactive environment. **Right:** Examples of movable object detection using SC-experiments and a mask predictor trained in the Placida environment. For each SC-experiment, we indicate the two resulting observations and one mask that highlights the object discovery. Note that the left prediction is performed in the Bolton environment, indicating that the method generalizes to unseen environments.

Do SC-type experiments allow detecting immovable objects? Results presented in Fig.1 show that the mask predictor is also able to accurately predict when the observations are different or identical. By isolating those two cases from the case where only one or a few objects have moved, we can compute a local SCP value that tells us whether the agent interacted with an immovable object during the SC-experiment. We can compute this local SCP value for different starting positions in the environment, and then construct a map of immovable objects in the environment. We present this map in Fig.2 for the arm's DOF that is closer to the body of the agent (we choose this DOF with the same reasoning as the previous result). Results show that regions with low local SCP value correspond to regions where there are walls and immovable objects in the way of Fetch's arm.

Indeed, in the kitchen part (room at the top), the space is cramped and so most of the positions indicate low SCP (less than 0.4) because of the interactions induced with the furniture. In the living room (main room) and the bedroom (at the left), most empty space show high local SCP (around 0.8 and 1.0). Notice how the local SCP value is also high around objects that are low and thus have to low chance to interact with the arm (bed in the bedroom, low table in the living room). We thus obtain a mapping of immovable objects in the environment using SCP.

Does our method generalizes to different environments?

In principle, this movable and immovable object detection method is designed to work in any environment, because it only relies on having a precise mask predictor, which we show can be achieved. In our experiments, we trained the mask predictor in the Placida environment and then verified that the model was still useful in the Rs-interactive and Bolton environments. In Fig.2, we show preliminary results that indicate that the mask predictor can indeed be used in other environments. It would require a finer analysis to assess how general our method is. This study of generalizability is not in the scope of this paper and left as future work.

Alternatives are not adapted. Alternatives to SC-experiments such as just playing an action sequence and comparing the first and last observations would detect much fewer objects because many experiments would result in a complete image change where the SC-experiments would highlight only a particular object. Another alternative would be to start in a position, play an action sequence, and then go back to this starting point and compare what's changed. While this approach would be comparable for movable object detection, this would not allow detecting immovable objects.

Conclusion. We showed how sensory commutativity can be used for movable and immovable object detection in the case of embodied agents.

References

[1] Hugo Caselles-Dupré, Michael Garcia-Ortiz, and David Filliat. Symmetry-based disentangled representation learning requires interaction with environments. In NeurIPS, 2019.

[2] Hugo Caselles-Dupré, Michael Garcia-Ortiz, and David Filliat. On the sensory commutativity of action sequences for embodied agents. arXiv preprint arXiv:2002.05630, 2020.

[3] Céline Craye, David Filliat, and Jean-François Goudou. Exploration strategies for incremental learning of object-based visual saliency. In 2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), pages 13–18. IEEE, 2015.

[4] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick Van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. arXiv preprint arXiv:1504.06852, 2015.

[5] Dibya Ghosh, Abhishek Gupta, and Sergey Levine. Learning actionable representations with goal-conditioned policies. arXiv preprint arXiv:1811.07819, 2018.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.

[7] Donald D Hoffman. The interface theory of perception. Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, 2:1–24, 2018.

[8] Scott P Johnson. How infants learn about the visual world. Cognitive Science, 34(7):1158–1184, 2010.

[9] Rico Jonschkowski and Austin Stone. Towards object detection from motion. arXiv preprint arXiv:1909.12950, 2019.

[10] Franz Kaufmann. Development of motion perception in early infancy. European Journal of Pediatrics, 154(4):S48–S53, 1995.

[11] Alban Laflaquière. Unsupervised emergence of spatial structure from sensorimotor prediction. arXiv preprint arXiv:1810.01344, 2018.

[12] Natalia Lyubova, Serena Ivaldi, and David Filliat. From passive to interactive object learning and recognition through self-identification on a humanoid robot. Autonomous Robots, page 23, 2015.

[13] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In Proceedings of the IEEE International Conference on Computer Vision, pages 9226–9235, 2019.

[14] Henri Poincaré. L'espace et la géométrie. 1895.

[15] Valentin Thomas, Jules Pondard, Emmanuel Bengio, Marc Sarfati, Philippe Beaudoin, Marie-Jean Meurs, Joelle Pineau, Doina Precup, and Yoshua Bengio. Independently controllable features. arXiv preprint arXiv:1708.01289, 2017.

[16] Melonee Wise, Michael Ferguson, Derek King, Eric Diehr, and David Dymesich. Fetch and freight: Standard platforms for service robot applications. 2016.

[17] Fei Xia, William B Shen, Chengshu Li, Priya Kasimbeg, Micael Edmond Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments. IEEE Robotics and Automation Letters, 5(2):713–720, 2020.

A Experimental details for object detection experiments

We provide further details on the object detection experiments:

- The dataset is composed of roughly 10k instances for each possible outcome (identical, completely different, identical up to moved objects).
- In order to generate the data for the "completely different" outcome, we apply a 90 degrees rotation to the robot.
- For the inference results on movable objects, we experimented with two strategies for the action sequence. Either 20 steps random action sequences or pre-determined action sequences (10 steps where the arm moves left, then 10 steps where the arm moves right).
- For the immovable object detection and creation of the map, we use random action sequences of 100 steps.