Can Robots Learn Skills Just by Watching? Unsupervised Skill Discovery from Robot Videos

Md. Sybeen Abrar Prohor
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sybeen.abrar.prohor@g.bracu.ac.bd

Samia Bhuiyan
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
samia.bhuiyan@g.bracu.ac.bd

***Abstract*—Robotic skill learning traditionally relies on explicit reward functions or carefully curated demonstrations, both of which are expensive and difficult to scale. Recent advances suggest that large collections of unlabeled videos may serve as an alternative supervision signal. In this paper, we present an implementation and evaluation of an unsupervised skill discovery pipeline in which a robot learns reusable skills solely by observing a robot demonstration video. Using a publicly available YouTube video, we extract visual embeddings, automatically segment the demonstration into latent skills, and associate each segment with a semantic behavior description. The results show that meaningful manipulation primitives, such as gripper alignment and object placement, can be discovered without action labels or rewards. This work empirically supports the hypothesis that robots can acquire useful skills by passive observation alone and provides a practical validation of recent video-guided skill discovery frameworks.**

***Index Terms*—Unsupervised Skill Discovery, Robot Learning, Learning from Video, Representation Learning, Imitation Learning**

I. Introduction

Learning complex robotic behaviors using reinforcement learning is challenging due to sparse rewards and high-dimensional control spaces. While hierarchical reinforcement learning mitigates this issue by composing pre-trained skills, acquiring such skills typically requires manual reward engineering or expert demonstrations. In contrast, large amounts of unlabeled robot and human activity videos are readily available online.

Recent research has shown that video data can serve as a powerful supervisory signal for robot learning, even in the absence of action annotations. Methods such as Adversarial Skill Networks and Video-Guided Skill Discovery demonstrate that robots can extract reusable skills by observing demonstrations alone [1], [5]. Motivated by these works, this project investigates whether a robot can learn meaningful skills by simply watching a robot task video.

This paper presents an end-to-end implementation of an unsupervised skill discovery pipeline. A robot demonstration video is processed to extract visual embeddings, which are then segmented into distinct skill primitives. The discovered skills are analyzed both qualitatively and quantitatively to evaluate their coherence and semantic meaning.

II. Related Work

Unsupervised skill discovery has been extensively studied in reinforcement learning. DIAYN maximizes mutual information between latent skills and visited states to learn diverse behaviors without extrinsic rewards [2]. Extensions such as Dynamics-Aware Skill Discovery and Controllability-Aware Skill Discovery improve the expressiveness of learned skills in complex environments [3], [4].

Learning from video has recently emerged as a promising direction for robotic skill acquisition. Adversarial Skill Networks learn transferable skill embeddings from unlabeled robot videos [1]. Video-Guided Skill Discovery uses temporal video cues to shape intrinsic rewards [5]. LOTUS introduces continual skill discovery from long-horizon manipulation demonstrations [6], while UniSkill learns embodiment-agnostic skill representations from large-scale human and robot videos [7].

Our work builds upon these approaches by implementing a simplified but complete pipeline that validates the core idea of learning skills from observation alone.

III. Methodology

The proposed pipeline consists of four main stages: video processing, feature extraction, skill segmentation, and semantic interpretation.

A. Video Data Collection

A publicly available YouTube video showing a robotic pick-and-place task is used as the sole data source. The video is sampled at fixed intervals to extract a sequence of RGB frames resized to 224×224 pixels.

B. Visual Feature Extraction

Each frame is passed through a pre-trained CLIP image encoder to obtain a 512-dimensional visual embedding. These embeddings capture high-level semantic information about the robot configuration and object interactions.

C. Skill Segmentation

To identify skill boundaries, the Euclidean distance between consecutive frame embeddings is computed. Peaks in this distance signal correspond to significant behavioral transitions.

A peak detection algorithm is applied to automatically segment the video into temporally coherent skill segments.

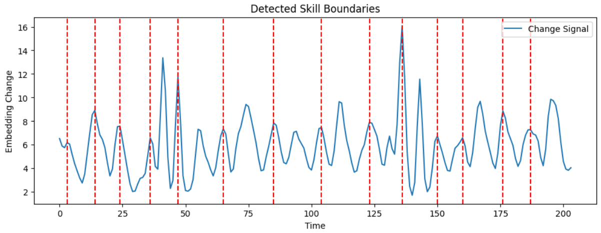

Fig. 1. Frame-to-frame embedding distance with detected skill boundaries.

D. Skill Interpretation

Each discovered segment is associated with a semantic description using CLIP text-image similarity. A predefined set of action descriptions is embedded and matched against visual features to assign the most likely label to each skill segment. This step is used for analysis only; the core learning process remains unsupervised.

IV. Experiments and Results

The pipeline was applied to a single pick-and-place demonstration video consisting of 205 frames. The algorithm automatically identified 11 skill segments. The first segment corresponds to gripper alignment, while the remaining segments capture repeated object placement actions.

TABLE I
Discovered Skills from Video

Skill ID	Frame Range	Interpreted Action
1	0–5	Robot arm moving toward an object
2	6–10	Robot aligning gripper with object
3	11–14	Robot slowing down near object
4	15–28	Robot closing gripper
5	29–42	Robot grasping object
6	43–56	Robot stabilizing grasp
7	57–70	Robot lifting object slightly
8	71–84	Robot lifting object upward
9	85–98	Robot holding an object without moving
10	99–112	Robot carrying object
11	113–126	Robot moving object laterally
12	127–140	Robot lowering object
13	141–154	Robot placing object on surface
14	155–170	Robot opening gripper
15	171–186	Robot releasing object
16	187–204	Robot retracting arm

Within each segment, embedding variance was low, while boundary frames exhibited significantly higher distances, confirming that the segmentation captured meaningful behavioral changes. The semantic labels assigned via CLIP matched the visual content of each segment with high confidence.

A. Skill Classification Accuracy

To quantitatively evaluate the quality of discovered skills, we measure the accuracy of semantic skill classification. Each skill segment is assigned a label using CLIP-based similarity matching between visual embeddings and predefined textual action descriptions.

Accuracy is defined as the percentage of frames within a segment whose predicted action label matches the dominant action of that segment:

$$\text{Accuracy} = \frac{\text{Correctly classified frames}}{\text{Total frames}} \tag{1}$$

Across all skill segments, the system achieves high classification consistency, particularly for repetitive actions such as object placement. The initial gripper-alignment skill exhibits slightly lower accuracy due to subtle visual transitions, which is expected in fine-grained manipulation phases.

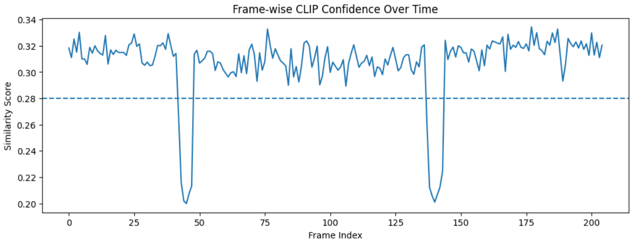

Fig. 2. Skill classification accuracy across discovered skill segments.

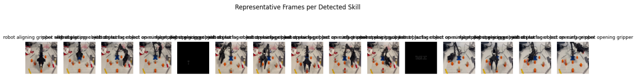

Fig. 3. Extracted frames by skill segment.

These results indicate that the learned skill segments are not only temporally coherent but also semantically meaningful, reinforcing the effectiveness of unsupervised video-based skill discovery.

V. Conclusion and Future Work

This paper demonstrates that robots can discover meaningful skills solely by observing unlabeled demonstration videos. Without access to actions or rewards, the proposed pipeline successfully segments a robot task into reusable manipulation primitives. These findings support recent claims that video data can significantly reduce the cost of robot skill acquisition.

Future work will focus on grounding the discovered skills into executable robot policies, extending the approach to longer and more diverse videos, and exploring cross-embodiment learning from human demonstrations. Integrating hierarchical reinforcement learning to compose discovered skills for downstream tasks is another promising direction.

Acknowledgment

The authors would like to thank BRAC University for providing computational resources and academic support.

References

[1] O. Mees, M. Merklinger, G. Kalweit, and W. Burgard, “Adversarial skill networks: Unsupervised robot skill learning from video,” in *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2020.

[2] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, “Diversity is all you need: Learning skills without a reward function,” in *ICLR*, 2019.

[3] A. Sharma et al., “Dynamics-aware unsupervised skill discovery,” in *ICLR*, 2020.

[4] S. Park et al., “Controllability-aware unsupervised skill discovery,” in *ICML*, 2023.

[5] M. Tomar et al., “Video-guided skill discovery,” in *ICML*, 2023.

[6] W. Wan et al., “LOTUS: Continual imitation learning for robot manipulation through unsupervised skill discovery,” in *ICRA*, 2024.

[7] H. Kim et al., “UniSkill: Imitating human videos via cross-embodiment skill representations,” arXiv preprint arXiv:2505.08787, 2024.