One-Shot Robot Manipulation from Human Video Demonstrations

Jitpuwapat Mokkamakkul

May 1, 2025

Advisor: George Konidaris — **Reader:** Srinath Sridhar

Table of Contents

1. Introduction

With the rapid development of AI systems, a lot of hope has been placed in the field of robotics to experience a similar jump in capability. Recent advances in machine learning, particularly in large-scale models for vision and language, have demonstrated that flexible, general-purpose systems are within reach for certain domains (Awais). However, robotics has yet to experience a comparable leap. Physical interaction with the real world — especially manipulation tasks in unstructured environments — remains a difficult and largely unsolved problem. Robots today typically require large amounts of task-specific demonstrations, supervised data, and carefully controlled settings to succeed at even simple manipulation tasks.

One bottleneck is the amount of methods that exist to allow robots to quickly and flexibly acquire new skills. One way we have looked into solving this problem is through giving robots demonstrations of tasks prior to doing a task. Traditional methods involve reinforcement learning initialized with a human physically moving the robot hand to do the task (Carfi). However, humans can learn new manipulation tasks by observing others, often needing only a single demonstration to understand the core principles. Hence, more modern methods of giving demonstrations to robots involve using advancements in computer vision and natural language processing to simply have the robot "watch" a human do a task and learn the necessary features required to do a task (Carfi). Building on this further, if robots could watch a human perform a task once and then successfully reproduce it, it would mark an important step toward building generalist robots that can adapt to new environments, objects, and goals with minimal intervention.

In this work, we explore the problem of enabling robots to perform manipulation tasks one-shot after observing a human demonstration, recorded through an RGB-D camera calibrated with the robot. By tracking the hand in the video and providing a text prompt describing the object being manipulated in the frame, we are able to extract the frame contact that occurs and the trajectory the hand takes through the manipulation. Then, we use motion planning to have the robot follow the trajectory extracted from the video. This approach attempts to dramatically reduce the amount of supervision and pre-training typically required for robotic learning, making it possible for end-users to teach robots new skills without technical expertise or access to large datasets.

Enabling one-shot visual imitation from human videos could have very important implications for the deployment of robots in real-world environments. It would open doors to household assistants that learn new chores on the fly, industrial robots that adapt to novel

workflows, and service robots that customize their behavior to individual users' needs. Ultimately, this work seeks to help bridge the gap between the impressive generalization abilities seen in AI systems and the currently narrow, task-specific capabilities of modern robots.

2. Background and Related Works

2.1. One-Shot Manipulation from RGB/RGB-D Demonstrations

Recent progress in one-shot manipulation from RGB/RGB-D demonstrations reflects a convergence of advancements in vision, action understanding, and affordance learning. Key methods often rely on extracting hand-object interactions from human videos, leveraging 3D pose estimation and egocentric datasets. One prominent approach is WHIRL, where FrankMocap (Rong, Shiratori, and Joo) is used to track human hands and remove the agent (human or robot) from frames, enabling the model to predict object motion through manipulation tasks (Bahl, Gupta, and Pathak). VideoDex, on the other hand, learns dexterous manipulation policies from Internet videos but focuses on mapping hand poses to robot end effector poses throughout interaction, demonstrating the scalability of learning from unstructured sources and the scale of features that can be extracted from RGB input (Shaw, Bahl, and Pathak).

Other research like Point Policy (Haldar and Pinto) additionally leverages RGB-D demonstrations to find keypoints on the human hand that can be mapped to the robot end effector, similarly to VideoDex, and is able to outperform previous baselines in various tasks. Finally, some research leverages multiple lab demonstrations that showcase multiple scenes or objects, allowing robots to generalize to new scenes and similar objects (Wang et al.).

2.2. Zero-Shot Manipulation from Human Videos

Zero-shot manipulation from human videos seeks to let robots perform unseen tasks by developing an intuition from training on large datasets. Unlike one-shot RGB-D demonstrations that often rely on controlled visual inputs, these methods emphasize learning affordances and skills from passive, often noisy, video data. For instance, Bharadhwaj et al. proposed methods where robots acquire manipulation capabilities by observing human-object interactions without any task labels, enabling flexible behavior transfer across categories (Bharadhwaj et al.).

Oftentimes, research in this field leverages large datasets such as EPICKitchens (Damen et al.), a labelled database with egocentric videos of humans interacting with objects in a kitchen, to train large models that can predict contact points and sometimes post contact trajectory given an object. Vision Robotics Bridge is one such example of a model built to predict contact points and a vector representing post contact trajectory (Bahl, Shikhar, et al.). Another model, Robo-ABC, only predicts contact points, and is also trained on EpicKitchens and performs well on pick and place tasks (when the motion after contact is simple).

Depth information can also now be leveraged from just RGB training data with tools like Depth Anything (Yang et al.), which extends the idea of large-scale pretraining to depth maps, enriching object and scene understanding in manipulation settings. One such model that utilizes modern depth estimation algorithms is Vidbot, which combines depth estimation and training on the EpicKitchens database to learn language conditioned 3D affordances (Chen). Dataset development specifically tailored to 3D manipulation has also accelerated progress. EPIC Fields (Tschernezki et al.) enriches egocentric video understanding by combining 3D reconstructions with temporal interaction data.

2.3. Hand and Object Detection

Accurate hand and object detection is foundational for learning from human demonstrations and building reliable databases like EpicKitchens, to enable models to localize interaction regions and detect contact. Recent research has significantly advanced the precision and scalability of detecting hands and objects in complex, real-world scenes. Modern approaches leverage both monocular RGB and RGB-D inputs, with increasing reliance on large-scale datasets and foundation models for generalization.

One stream of work focuses on reconstructing detailed hand poses. Systems like FrankMocap (Rong, Shiratori, and Joo) offer full-body and hand pose estimation from monocular RGB, making them widely adopted in manipulation research due to their accessibility and efficiency. More recent methods leverage transformers for higher fidelity hand tracking in 3D, such Hamer, proposed by Pavlakos et al., which enhances hand reconstruction even under occlusion or poor lighting conditions. Shan, Geng, Shu, and Fouhey proposed a foundational model for detecting hands in contact and whether contact is happening, uses overlapping bounding boxes as a method for determining contact, a method which is also leveraged in this thesis. This model also provides a baseline for object detection.

However, more modern object detection models utilize advances in large language models to scope which objects in the frame are relevant. Of particular relevance is GroundingDINO, which outputs bounding boxes for objects in a monocular RGB image given a text prompt (Liu, Tripathi, Majumdar, and Wang).

3. Methodology/Design

3.1. Hand and Object Bounding

State of the art models were used for hand and object detection and integrated into a single pipeline. Hamer (Pavlakos et al.) was used as the hand model, and outputs an array of hand keypoints for both the right and left hand. The minimum and maximum width and length values were stored to create a bounding box around each hand. GroundingDINO (Liu) was used to detect objects in the frame. If the bounding boxes of a hand and object overlapped past a threshold of 0.3, then the hand was said to be in contact.

For each frame, hand and object bounding boxes were detected; and since some frames may be unclear, there were outlier frames where contact was detected when there was no contact. To avoid this, a rolling median outlier detection algorithm was added to ensure only prolonged periods of detected contact are stored.

Throughout the contact frames, the center of the hand bounding box was stored as the point where the hand was when contact was occuring. These points were concacated to create a trajectory in pixel space throughout contact with the object.

3.2. Mapping Camera Coordinates to World Coordinates

Once pixel-space frames for the hand during contact are extracted, their depth values are found and mapped to the camera frame using collected camera intrinsics. The intrinsic matrix includes focal length and the location of the principal point (Tsai). The calculation for this looks like the following:

$$P_c = Z \cdot K^{-1} \cdot I$$

P_C — Camera coordinates

Z — Depth coordinates

K — Intrinsic matrix

I — Image coordinates

Furthermore, these points in the camera frame were mapped to the environment world frame using extrinsics (rotation and translation matrixes):

$$P_w = R \cdot P_c + t$$

P_w — World coordinates

R — Rotation

t — Translation

3.3. Motion Planning to the Extracted Points

Using the list of world-frame points of the hand in contact, MoveIt (Șucan) was used to motion plan a robot arm to the positions given. A quarternion pose corresponding to the robot end effector pointing directly downwards was also used. At the first point in the list (the point where contact first occurs), the end effector was manually set to close, and at the last point in the list, the end effector was set to open.

3.4. Experimental Setup

Figure 1 and 2: Experimental setup for two tasks displaying multisense camera, Kuka iiwa arm, BarrettHand end effector; with the box opening task on the left and the pick and place task on the right.

Experiments were ran with a setup including a Kuka iiwa arm, BarrettHand end effector, and a calibrated Multisense camera (using which video demonstrations were also taken). Two tasks were demonstrated for the robot to follow:

1. A pick and place task with a green stuffed toy
2. A box opening task with a box containing a large handle

These tasks were selected to test both basic object displacement and a more complex manipulation task involving object affordances and non-linear motion.

For each of these tasks, 5 videos were taken of a human interacting with the object and robot poses were extracted using methods explained in the above sections. Then, MoveIt was used to have the robot perform the given task for each of the trajectories extracted from each video. Sucess in each trial was documented and the next section shows results of the average success rate accross the five trials.

For the pick and place task, success was defined as the robot getting a stable grasp on the object and being able to lift the object following the trajectory in the video, without dropping the object before the end effector is opened.

For the box opening task, success was defined as the robot getting a stable grasp on the handle of the box and being able to open the box at least 90 degrees.

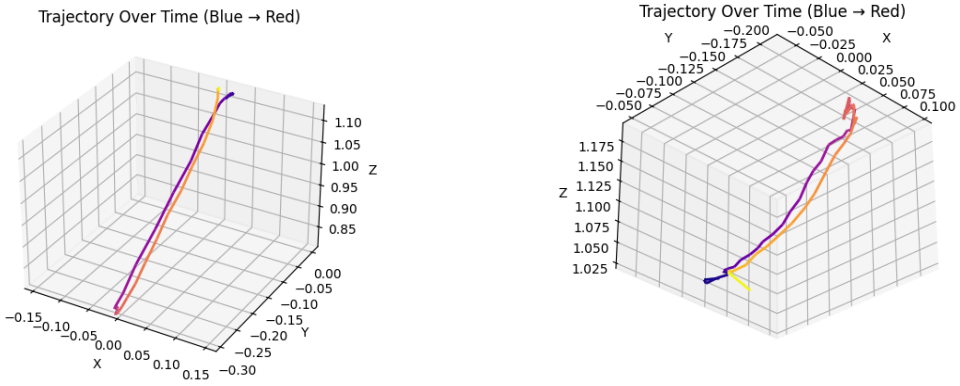
4. Results

4.1. Results for Both Tasks

Table 1: Average Success Rate for Each Task Over Five Trials

Task	Average Success (over 5 trials)
Pick and Place	1.0
Box Opening	0.6

Figures 3 and 4: Trajectory During Contact Extracted From Human Demonstration in Camera Frame (Left: Pick and Place, Right: Box Opening)



X and Y are the width and height values relative to the camera position and Z is the depth, or direct distance from the camera. The trajectories are at an angle because the multisense was mounted at an angle from the objects.

Results indicate a performance gap between the two tasks (pick and place + box opening).

The pick-and-place task achieved a 100% success rate, with the robot reliably identifying the correct grasp location and replicating the trajectory from the human demonstration with high precision.

However, the box opening task achieved a 60% success rate, completing the task in 3 out of 5 trials. This disparity shows the difficulty posed by manipulating objects with joint constraints, where precision in contact initiation and trajectory play a much more important role than with the pick and place task with a malleable object.

At the same time, the trajectories extracted for both tasks look very smooth and not too noisy, suggesting that the system can reliably extract trajectories, but face challenges when dealing with more complex tasks or objects that are more tightly constrained.

4.2. Qualitative Analysis and Failure Modes

To better understand the causes of failure in the box opening task, qualitative data was collected. The two rounds in which the box opening task failed were due to the following failure modes:

1. Missing the contact point - In one case, the robot was not able to grasp the handle properly, potentially due to pose misalignment or slight errors in translating human to robot poses, and thus couldn't manipulate the object
2. Pushing the entire box - In another failed trial, instead of opening the lid, the robot pushed the entire box forewards. This potentially occured because the contact was initiated too low and because the robot wasn't given any understanding of the object.

In addition to these task-specific failures, some systemic limitations were also noticed. In some instances, the robot would plan potentially dangerous or aggressive trajectories to get to the initial contact point. In others, the poses given to the robot would cause it to exceed its joint limits and thus, poses needed to be manually edited or videos needed to be retaken.

5. Discussion

5.1. Summary of Main Findings

Through this work, I've demonstrated a system for one-shot manipulation learning from RGB-D video, where a robot performs a manipulation task after observing a single human demonstration. Two tasks are evaluated, namely: (1) a pick-and-place task involving a plush toy, and (2) an articulated interaction task involving the opening and closing of a box.

The pipeline was able to achieve 100% success on the pick-and-place task, demonstrating the reliability of mapping extracted video features to practical robot trajectories. On the box task, which requires more detailed movement, the system achieves 60% success.

Compared to prior approaches such as WHIRL, which achieves 60% and 73% success on drawer and door manipulation tasks respectively, this method demonstrates competitive performance despite using only a single RGB-D demonstration per task. Additionally while WHIRL relies on additional training using reinforcement learning, our approach eliminates the need for such exploration and learning, highlighting the potential of using extracted video features alone as a more efficient and scalable alternative for teaching robots manipulation skills. On the other hand, this also demonstrates the potential of this method to perform even better than existing baselines if leveraging learning-based algorithms at interaction time.

To better understand the robot's behavior, we also analyze and visualize the extracted 3D motion trajectories used to replicate the demonstrations. These diagrams reveal that in both tasks, trajectories are consistent and closely follow the demonstrated motion path. However, the occassional failures in the box opening task are likely not from the trajectory extraction, but from the lack of robustness of motion planning and the methods used for mapping hand poses to end effector poses in this thesis. For instance, if the center of the hand is slightly higher or lower than the origin of the end effector, then then this can affect results for more fine-grained tasks like the box opening.

Overall, these results support the feasibility of using RGB-D video for efficient robot manipulation, even without in real life training, and point toward a promising direction for building generalist manipulation systems capable of learning diverse tasks from human videos.

5.2. Limitations of Methods Used

While the proposed one-shot manipulation system demonstrates promising results, several limitations in the current methodology constrain its generality and performance,

particularly in more complex tasks.

As noted in the discussion of the box opening results, the occasional failures in this task are likely not due to errors in the extracted demonstration trajectories, but rather stem from the lack of robustness in motion planning and the hand-to-end-effector mapping strategy used in this thesis. The approach assumes a relatively rigid alignment between the human demonstrator's hand poses and the robot's end-effector poses. However, if the center of the human hand is slightly higher or lower than the assumed origin of the end effector, even minor discrepancies can introduce significant errors — especially in fine-grained tasks like opening a box, where precise contact and motion trajectories are required. These sensitivities highlight a need for either better calibration or a method to robustly map hand poses to end effector poses.

This work uses RGB-D video to capture both color and depth information for trajectory extraction, which provides accurate depth information necessary for precise manipulation. However, this reliance also introduces a potential limitation. In contrast to the vast availability of RGB video data online, on Youtube or egocentric datasets like EpicKitchens, depth data is relatively scarce. This restricts the scalability of the method if it were to be integrated into larger systems trained on diverse human demonstrations from the web.

Finally, the current system relies primarily on deterministic motion planning rather than incorporating a learning-based execution policy. While this simplifies the system and provides more interpretability, it also limits the robot's ability to adapt to small errors or changes in the environment (demonstrated in the failure modes discussed). In tasks with more complex object dynamics or requiring fine-grained movements— such as the box opening task — even small discrepancies in pose estimation or contact timing can lead to task failure. A learned controller, trained to correct or compensate for such errors based on prior experience, may be better suited to handle these cases. Integrating learning into the execution stage could improve robustness and potentially allow for more flexible generalization across object types and task variations.

5.3. Further Work

To address some limitations mentioned, many promising directions can be pursued to improve robustness, generalization, and scalability of manipulation learning from video.

A key limitation in the current pipeline is the simplistic mapping from human hand poses to robot end-effector poses. Future work should incorporate more advanced hand pose tracking throughout the video and explore mapping such poses to equivelant robot poses. With a reliable

method to map human to robot poses in the object frame, it may be possible to generalize across similar objects with larger or smaller components.

To remove dependence on RGB-D sensors, future systems could replace depth cameras with monocular depth estimation from standard RGB video. Since we typically are concerned with the actual moment of manipulation and the points during manipulation, relative depth mapped to the object frame is likely sufficient. This would open the door to leveraging vast amounts of existing RGB-only video data and potentially enable the use of large-scale pretraining or foundation models based on everyday human demonstrations found online.

Lastly, a lot of prior work makes use of the EpicKitchens dataset, as it is the only large dataset that contains labelled annotations for everyday object interactions. However, a lot of models that EpicKitchens used to annotate their data are now outdated and there are some labels that dictate an interaction is occuring, but it is not shown given the angle of the camera. Hence, for further development of manipulation-based robot algorithms, it may be useful to create new datasets similar to EpicKitchens that have more robust labels and reliable annotations.

Acknowledgements

Thank you to my wonderful advisor, George Konidaris, who was kind enough to let a second semester freshman who knew nothing about AI and robotics to sit in on lab meetings. He has been nothing but patient when I've hit walls in my research or when I've had to take time to learn new skills, such as ROS and camera intrinsics/extrinsics. Without support from Professor Konidaris and his kind and soothing personality, I would not have been able to accomplish all this within my time at Brown.

Next, thank you to everyone in the lab that has helped me with this project. Most of the people I've interacted with are in the Brown Intelligent Robot Lab, but I've also received support from people in other robotics and Computer Vision labs at Brown. In particular, thank you Shivam Vats, Varun Satheesh, Sudarshan Haritas, and Tao Lu for being wonderful people to work with or bounce ideas off of when building out this project. Without Shivam's experience and guidance, I would not have been able to find the necessary resources and papers to complete this research.

Also, thank you to all the people that have guided me through the earlier stages of this project, where I experienced many roadblocks, and still supported me despite it. Without all of you, I would not have been able to make it to where I am today. Namely, Skye Thompson, Ben Abbatematteo, Rahul Sajnani, Seiji Shaw, and Eric Rosen.

Lastly, special shoutout to those I got to know better in the lab over the summer of 2023. That summer was one of my most memorable experiences at Brown thanks to all the wonderful people I got to know in the robotics lab.

Bibliography

Abbatematteo, Ben, et al. "Composable Interaction Primitives: A Structured Policy Class for Efficiently Learning Sustained-Contact Manipulation Skills." 2024 IEEE International Conference on Robotics and Automation (ICRA), 2024, pp. 7522–7529. doi:10.1109/ICRA57147.2024.10610846.

Awais, Muhammad, et al. "Foundation Models Defining a New Era in Vision: A Survey and Outlook." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 47, no. 4, 2025, pp. 2245–2264. doi:10.1109/TPAMI.2024.3506283.

Bahl, Shikhar, Abhinav Gupta, and Deepak Pathak. Human-to-Robot Imitation in the Wild. 2022, arXiv, https://arxiv.org/abs/2207.09450.

Bahl, Shikhar, et al. Affordances from Human Videos as a Versatile Representation for Robotics. 2023, arXiv, https://arxiv.org/abs/2304.08488.

Bharadhwaj, Homanga, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Zero-Shot Robot Manipulation from Passive Human Videos. 2023, arXiv, https://arxiv.org/abs/2302.02011.

Carfì, Alessandro, et al. "Hand-Object Interaction: From Human Demonstrations to Robot Manipulation." Frontiers in Robotics and AI, vol. 8, 2021, article 714023. doi:10.3389/frobt.2021.714023.

Chen, Hanzhi, et al. VidBot: Learning Generalizable 3D Actions from In-the-Wild 2D Human Videos for Zero-Shot Robotic Manipulation. 2025, arXiv, https://arxiv.org/abs/2503.07135.

Damen, Dima, et al. "Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100." International Journal of Computer Vision (IJCV), vol. 130, 2022, pp. 33–55, https://doi.org/10.1007/s11263-021-01531-2.

Haldar, Siddhant, and Lerrel Pinto. Point Policy: Unifying Observations and Actions with Key Points for Robot Manipulation. 2025, arXiv, https://arxiv.org/abs/2502.20391.

Ju, Yuanchen, et al. "Robo-abc: Affordance Generalization Beyond Categories via Semantic Correspondence for Robot Manipulation." European Conference on Computer Vision, 2025, Springer, pp. 222–239.

Kuang, Yuxuan, et al. RAM: Retrieval-Based Affordance Transfer for Generalizable Zero-Shot Robotic Manipulation. 2024, arXiv, https://arxiv.org/abs/2407.04689.

Liu, Shaowei, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint Hand Motion and Interaction Hotspots Prediction from Egocentric Videos. 2022, arXiv, https://arxiv.org/abs/2204.01696.

Liu, Shilong, et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. 2024, arXiv, https://arxiv.org/abs/2303.05499.

Mendonca, Russell, Shikhar Bahl, and Deepak Pathak. ALAN: Autonomously Exploring Robotic Agents in the Real World. 2023, arXiv, https://arxiv.org/abs/2302.06604.

Niu, Dantong, et al. Pre-training Auto-regressive Robotic Models with 4D Representations. 2025, arXiv, https://arxiv.org/abs/2502.13142.

Pavlakos, Georgios, et al. Reconstructing Hands in 3D with Transformers. 2023, arXiv, https://arxiv.org/abs/2312.05251.

Rong, Yu, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. 2021, arXiv, https://arxiv.org/abs/2108.06428.

Shan, Dandan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding Human Hands in Contact at Internet Scale. 2020, arXiv, https://arxiv.org/abs/2006.06669.

Shaw, Kenneth, Shikhar Bahl, and Deepak Pathak. VideoDex: Learning Dexterity from Internet Videos. 2022, arXiv, https://arxiv.org/abs/2212.04498.

Şucan, Ioan A., and Sachin Chitta. "MoveIt." MoveIt, [Online] Available at moveit.ros.org.

Tsai, Roger Y. *A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using off-the-Shelf TV Cameras and Lenses*. no. 4, Aug. 1987, pp. 323–44, https://doi.org/10.1109/JRA.1987.1087109.

Tschernezki, Vadim, et al. "EPIC Fields: Marrying 3D Geometry and Video Understanding." Proceedings of the Neural Information Processing Systems (NeurIPS), 2023.

Wang, Jianren, et al. One-shot Video Imitation via Parameterized Symbolic Abstraction Graphs. 2024, arXiv, https://arxiv.org/abs/2408.12674.

Yang, Lihe, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. 2024, arXiv, https://arxiv.org/abs/2401.10891.