Semantic Understanding of Task Outcomes: Visually Identifying Failure Modes Autonomously Discovered in Simulation

Joseph Bowkett*, Joel Burdick, Larry Matthies, and Renaud Detry

***Abstract*— We present a model for identifying and recognizing task success and distinct modes of task failure in robot manipulation applications. Our model leverages physics simulation and clustering to learn symbolic failure modes, and a deep network to extract visual signatures for each mode and to guide failure recovery. We present an early experiment where we apply our model to the archetypal manipulation task of placing objects into a container. A CNN is trained on synthetic depth images generated and labeled in simulation, and we demonstrate the ability of the network to compute task outcomes in both synthetic and real depth images.**

I. Task Outcome Classification

Fallibility is not the sole preserve of mankind. Despite our efforts to make released systems function with absolute consistency, deployed robots can and do make mistakes.

As our field pushes further into unstructured environments, such as human workplaces and homes with increasingly generalized use cases, the incidence of failures is set to increase. For robots to be able to operate effectively in these environments, they must possess the ability to identify and correct any failures in tasks they are set, be these due to insufficient planning data, unforeseen impediments, or adversarial interference.

Over the past three decades, our community has constructed a solid understanding of the geometric aspects of manipulation – motion planning, grasp (hand/wrist pose) planning, manipulation control. By contrast, the *semantic* aspect of manipulation remains poorly understood. Concepts related to task success generalize poorly under the strictly geometric metrics that we currently use.

We propose a semantic task outcome model that leverages contact/physics simulation to parse the structure of a given behavioral domain and to extract a symbolic characterization of the nature of possible failures (or *failure modes* of the task). In turn, our model leverages an image classifier to capture the sensory context of a manipulation task, and to ground failure modes in perceptual data.

We propose to identify the failure modes of a given task by executing randomly-perturbed variants of reference trajectories provided by an instructor, and grouping those executions according to proximity in a space consisting of geometric measurement effected on end-of-task scene configurations. This work is conducted in simulation, which facilitates execution and exposes geometric parameters whose computation in the real world would be prohibitively time-consuming.

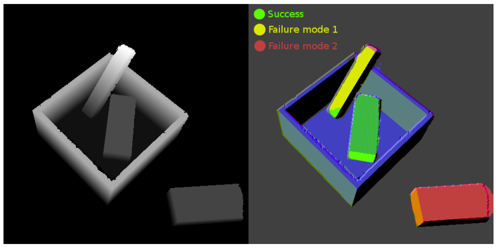

Fig. 1. Left: Greyscale render of synthetic depth image. Right: Task outcome labels provided by our model: green corresponds to a successfully-executed insertion task, yellow and red correspond to two modes of failure.

The responsibility of the image classifier is to identify whether a given perceptual representation of a scene indicates success or failure of the task, and in the case of failure, identify a failure mode. When a failure occurs, it can be taken as evidence that the information used to plan the task was either flawed or incomplete. For this reason, determining the outcome of general tasks in unstructured environments may benefit greatly from an unbiased assessor that ignores any *a priori* knowledge of the workspace. Accordingly, we implement task outcome classification with a convolutional neural network – a model known for its capacity to capture high-variance environmental parameters.

As an example, let us consider the task consists of inserting an object in a box shaped container by dropping the object from above. Different outcomes can be exposed by varying the pose from which the object is dropped, which could yield for instance *object in box* (success) *object fell outside of the box* (failure mode 1), and *object in the box but sticking out* (failure mode 2). These three labels naturally emerge by clustering the outcomes according to the distance between the object and the center of the box. The role of the image classifier amounts to capturing a direct mapping between an image of the end-of-task scene, and its label (success, or failure mode in case of a failure).

In prior art, much of the existing work on analysis of manipulation task outcomes focused on prediction, so as to minimize the chances of failure [1], [2]. While this greatly improves the likelihood of task success on the first attempt, we believe it necessary to consider the failures that will inevitably occur in these ever more general environments. Similar in nature to our work, Hanheide et al. described unexpected failures in motion planning as a mismatch be-

The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0016. ©2018 California Institute of Technology. Government sponsorship acknowledged.

*Corresponding author: jbowkett@caltech.edu

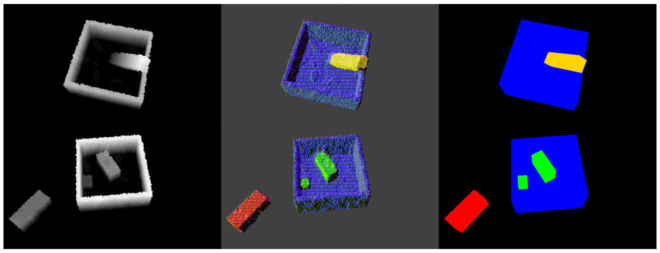

Fig. 2. Left: Synthetic depth images (with Kinect noise model). Center: Labels produced by task outcome model. Right: Ground truth labels.

tween expectation and experience [3]. Our work goes beyond Hanheide et al. by relaxing assumptions on the environment. Visual task success verification was investigated by Erkent et al. [4] by checking for completion while using visual servoing on various tasks. However, the authors do not attempt to classify types of failure if success is not detected in a given time. Saran et al. explored viewpoint selection for visually determining binary task failure [5], which is complementary to the work described here.

II. Object placement experiment

The archetypal manipulation task chosen for demonstrating the model is the object-container insert alluded to above. Placement of each object was deemed a success if the entire object was contained within the receiving box; any other result was deemed a failure, as seen in Fig. 1.

To identify failure modes, we simulated 10,000 variants of a reference insert scenario of which 95% are used for training and 5% for validation. We released between one and four box-shaped objects 1.5m above the center of a larger container of side length 1m (Fig. 1). The initial position of the objects was offset in a plane parallel to the ground, by a vector drawn randomly from a uniform distribution defined on $\{(x, y) : x, y \in [-60, 60]\}$. We simulated the perturbed episodes in Blender with the Bullet physics engine. We identified failure modes by clustering all object drops according to the distance between the center of the container and the center of mass of the dropped object in its final configuration. Clustering was implemented with DBSCAN [6], which yielded three clusters that correspond to the three outcomes that a human observer would intuitively identify, namely *object in container* (success), *object fell next to container* (failure mode 1), and *object in the container but sticking out (including object sitting on edge of container)* (failure mode 2).

To enable the identification of the three outcomes listed above in a new scene, we trained a CNN on a dataset consisting of labeled depth images. We captured one depth image of the final configuration of each of the episodes generated above using a realistic depth-camera sensor model [7], and we labeled all images with ground-truth outcomes yielded by the clustering algorithm. While the problem discussed above can be identified as a classification task, we opted to model both the outcome of the task and the spatial structure of the scene, to facilitate the definition of recovery actions in future work. Accordingly, instead of labeling each image, we labeled each pixel of each image according to the pixel's correspondence to (1) a successfully-inserted object, (2) an improperly-inserted object (failure mode 1), (3) an improperly-inserted object (failure mode 2), (4) the container, and (5) the background. We trained the fully-convolutional MultiNet architecture proposed by Teichmann et al. [8], [9] on this dataset. To measure the network's performance, we evaluated the network's ability to predict the presence or absence of each of the three outcome classes (success, failure mode 1, failure mode 2), by comparing the number of pixel labels belonging to each class to a hard-coded threshold. Our results showed that in this canonical experiment, the network identified the correct presence or absence of all outcomes in 86.8% of images in the synthetic validation set. Fig. 2 shows several depth images along with ground-truth and predicted labels. The first row of Fig. 2 shows an example of failure mode 1: the object is not entirely within the volume of the container. The second row shows two successful inserts (in green), and one failure of mode 2 (in red). Fig. 3 demonstrates the network's ability to predict outcome labels on real depth images from a Kinect camera.

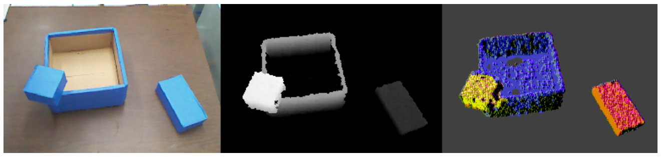

Fig. 3. Left: RGB image of real scene from Kinect camera. Center: Grayscale render of processed depth image from Kinect. Right: Overlay of segmented classes produced by CNN.

References

[1] P. Pastor, M. Kalakrishnan, S. Chitta, E. Theodorou, and S. Schaal, "Skill learning and task outcome prediction for manipulation," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011.

[2] B. Moldovan, P. Moreno, M. van Otterlo, J. Santos-Victor, and L. De Raedt, "Learning relational affordance models for robots in multi-object manipulation tasks," in *ICRA, 2012*. IEEE, 2012.

[3] M. Hanheide, M. Gobelbecker, G. S. Horn, A. Pronobis, K. Sjoo, A. Aydemir, P. Jensfelt, C. Gretton, R. Dearden, M. Janicek, H. Zender, G.-J. Kruijff, N. Hawes, and J. Wyatt, "Robot task planning and explanation in open and uncertain worlds," *Artificial Intelligence*, August 2015.

[4] O. Erkent, D. Shukla, and J. Piater, "Visual task outcome verification using deep learning," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept 2017.

[5] A. Saran, B. Lakic, S. Majumdar, J. Hess, and S. Niekum, "Viewpoint selection for visual failure detection," in *IROS, 2017*. IEEE, 2017.

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise," in *Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996.

[7] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree, "Blensor: Blender sensor simulation toolbox," in *Advances in Visual Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[8] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, "Multinet: Real-time joint semantic reasoning for autonomous driving," *arXiv preprint arXiv:1612.07695*, 2016.

[9] R. Detry, J. Papon, and L. Matthies, "Task-oriented grasping with semantic and geometric scene understanding," in *IROS*, 2017.