An RGBD segmentation model for robot vision learned from synthetic data

Jonathan C. Balloch and Sonia Chernova
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: balloch@gatech.edu and chernova@cc.gatech.edu

Abstract—The ability to segment objects for semantic reasoning about a complex scene is fundamental to interactive robot behaviors. However, most robot systems can only operate in unrealistic, structured environments. For robots to operate in unstructured, unknown environments and reason about interacting with the objects and agents of those environments, robots need to be able to segment the scene that they perceive. In this work, we examine the use of synthetic data to supplement training of complex deep models that do not on their own have enough labeled real-world training data. We train a state-of-the-art deep neural network designed for scene segmentation on different combinations of the real and synthetic data and show our model being able to segment video frames from a robot in spite of the model not having been trained on robot data.

I. INTRODUCTION

Accurate perception of objects is crucial for fundamental robot behaviors such as object interaction and manipulation, semantic mapping, and obstacle avoidance. In robotics research, robots usually depend on large amounts of prior knowledge about the environment. This could be an accurate database of 3D models of objects, making the objects and environment highly structured and easily distinguishable from each other, or are augmenting the environment with fiducials to enable detection and tracking. In practice, however, these approaches limit the ability of robots to function outside their research environments. For a robot to physically interact with objects and agents in an environment that is not a highly controlled space like a factory or laboratory, it would need to be able to see and reason about the unstructured and unknown composition of that environment. For this reason, semantic segmentation of scenes is of the utmost importance in robotics.

Fully convolutional neural networks (FCNs) have been demonstrated by the computer vision community to yield very promising results on learning the semantic object segmentation of images [13]. However, image dataset bias can make these datasets not representative of the data from robot sensors in the real world. Issues like camera shake, poor white and color balance, and rapid movement of both the robot and objects in the scene are rarely captured by current vision datasets, even those that attempt to make the training data more realistic or difficult by adding noise. As a result, these datasets can be insufficient for training a model to be used to segment real world video. In addition, there are only a few video datasets annotated for object or scene segmentation, and while some image datasets for other tasks have upwards of 1 million annotated images, the most comprehensive video segmentation datasets have only tens of thousands of annotated frames, which remains a major performance bottleneck in training these models. Canonical literature in data-driven approaches to machine learning often reiterate the “Rule of 10,” stating that for each class that the algorithm is attempting to learn there should be ten times as many training examples as there are degrees of freedom [9]. It is not precisely clear how many degrees of freedom are being learned in the semantic segmentation of images, as scaling and perspective represent variably more than two degrees of freedom each. However, it is clear that for popular semantic segmentation datasets like SUN3D and NYU RGB-D v2, where there are 5285 labeled frames for 37 classes and 1449 labeled frames for 13 classes respectively, there are not enough data annotated for a well-regularized representation of those classes, even if the classes were evenly distributed. Part of the reason for this is because both collecting and hand annotating object segments in each frame of a video dataset is extremely time intensive, especially compared to annotation of image classification and detection datasets. One possible solution to the problem of needing annotated data is to use synthetic data generated from accurate simulated environments to help train these models.

The use of deeply learned computer vision models for object segmentation on robot platforms is still relatively uncommon. This is in part because these models are computationally expensive to both test and train, but also largely because the robotics community is inherently focused on video, which has made less progress in segmentation than images until recently. We believe that by training a deep robot vision system initially on sufficiently large amounts of realistic synthetic data, a robot would be able to use such a vision system to effectively reason about unknown environments. The use of synthetic data to improve computer vision models is relevant to a variety of applications, but if successful this approach is uniquely well suited to improve vision in robotics because of the widespread use of realistic robotic simulators such as Gazebo, Klamp’t, OpenRAVE, and numerous proprietary simulators for specific robots. With our approach robots could actively generate labeled samples from simulations of their own environments to improve the quality of their vision systems. We show preliminary results that suggest that synthetic data generated
from simulation and automatically annotated can improve
semantic video segmentation enough to realistically deploy
these methods on robots.

II. RELATED WORK

Segmentation has gained much ground in the last few years
with developments of FCNs [13]. The approach of Chen et al.
using this fully convolutional structure with dense conditional
random fields, called Deeplab, is the underlying approach to
most of the state-of-the-art segmentation today [2]. Deeplab
is the underlying algorithm to the algorithm we use in this
research as well, and we expand on its capabilities to make it
more well suited to the challenges of robot vision.

Many popular segmentation datasets exist, including Pascal
VOC 2012, MS-COCO, KITTI driving dataset, and more
recently the CityScapes autonomous driving dataset in the
RGB space, the NYU v2 and SUN3D datasets in the RGB-
D space, and the DAVIS dataset for video segmentation [4]
[12] [3] [16] [24] [17]. Most of these datasets have a
predetermined set of classes, and pixels are all labeled as one
of those classes, or if they do not fit in any, they are not
labeled. For robots operating in uncontrolled environments,
however, it is not simply enough to segment a few objects, it is
often important to be able to segment the entire scene. Within
the computer vision community, the task of scene parsing,
or segmentation with a label for every pixel in the scene, is
most similar to our goal. This is the focus of the recent MIT
ADE20K dataset [27], which provides over twenty thousand
frames of both indoor and outdoor scenes with every pixel
labeled with one of 150 class labels.

There has also been recent similar work in using deep neural
networks for semantic segmentation of objects and scenes for
robotic systems, especially with regards to autonomous driving
[11]. Most of these efforts take advantage of semantics to
improve their segmentations [25] [23]. Song et al. ’s approach
focuses on attempting to perfectly segment a new space upon
first seeing it, and then operate on the assumption that the
object set will not grow much beyond the initial observation of
the environment [22]. This method is an important step in the
correct direction, but like prior less comprehensive methods
it depends on synthesizing a virtual map of the environment
and segmenting that map as opposed to explicitly segmenting
each frame. In a similar vein, McCormic et al. attempts to
improve semantic segmentation for the task of RGB-D SLAM
for robots [14]. This work is distinct in that we are motivated
to improve the frame-by-frame segmentation in robot vision
instead.

Synthetic data has been used to benchmark algorithm per-
formance, especially in the optical flow community where the
MPI-Sintel and Flying Chairs datasets are currently two of
the most heavily used benchmarks [1]. Only very recently
have researchers shown exciting results by training models
on synthetic segmentation data [18]. That work by Richter
et al. showed success on improving segmentation of real
images with synthetic RGB image data, however still only
produces tens of thousand of images, does not use video,
and does not have depth data, which is important for the
application of such datasets to robotics. To address this greater
call for large synthetic datasets, the synthetic autonomous
driving Synthia Dataset was released, which features over 200,000 high definition frames from video clips in a variety of different environments and weather patterns, annotated with 13-class segmentation labels and simulated data from 8 RGB cameras forming a binocular 360 camera, and 8 depth sensors [19]. Most similar to our research, Handa et al. created a comprehensive dataset synthetic annotated data from video captures of 3D simulated scenes [7]. They showed that while using only synthetic data did not alone approach the state-of-the-art, first training on synthetic depth data and then fine tuning with regular training data improved on the state-of-the-art models’ results for the NYU v2 and SUN3D datasets [6].

III. METHODS

Our approach has three components: the training data from synthetic and real image segmentation datasets, a vision-enabled robot designed for interaction, and a deep neural network model designed for scene segmenting. Our process leverages a current state-of-the-art segmentation model by training it on a large quantity of synthetic data, fine-tuning this model on annotated real world data, and then showing results on data captured from the robot.

For our synthetic training data, we utilize the recently released SceneNet RGB-D dataset [15]. This is the only dataset that has a very large amount of video data of diverse indoor scenes, includes depth data, and has semantic and instance segmentation annotations for every frame. The scenes are rendered very realistically, and data diversity is improved by assigning each simulation random lighting, camera trajectories, and textures, and adding extra noise to the final rendering. These techniques help prevent learning models from overfitting to any unnatural features of the data.

We use a Vector human robot interaction (HRI) robot developed by Stanley Innovation with a Kinect v2 for our experiments. We chose to use data from the Vector robot because the Kinect v2 sensor is a common vision sensor on robots and HRI robots are designed for interaction. As such, the data from this vision system is representative of robots designed to interact with their environment. The visual data we test on from this robot was captured raw using the Robot Operating System (ROS) from navigation around a space set up like a typical studio apartment, sometimes with human subjects, sometimes without. These captures were densely annotated using the iSeg video segmentation tool for evaluation [21]. The labels for all of these objects overlap with the ADE20K dataset labels.

The neural network model we use is the Pyramid Scene Parsing Network model from Zhao et al. (PSPNet) [26]. This model extends the Deeplab segmentation approach by using a hierarchical global pooling layer at the end of the network to prevent the loss of context in features and improve scale invariance. It is appropriate for our problem as it produced the best result on the ImageNet Scene Parsing Challenge over the MIT ADE20K scene parsing dataset [20] [27]. Starting with a
ResNet-50-based PSPNet model pre-trained on the ImageNet dataset, we first train our architecture, as shown in Figure 2, on the SceneNet RGB-D dataset, lowering the learning rate every 10000 iterations using the Caffe deep learning framework and a NVidia GTX 1060 GPU [10]. Then we fine-tune the network using the ADE20K dataset, as it has a larger and more expressive set of semantic classes that overlap with the real robot data. The result is a model trained on far more data, but with the same amount of real data and the same number of classes in the final model. This model is then tested on robot data which we annotated using a subset of the semantic classes used by ADE20K.

IV. EVALUATION AND RESULTS

We have been able to show that training this network on synthetic data can be used to obtain plausible segmentation results on real images such as that of the ADE20K dataset. In Figure 3, we show the segmentation results of our model trained on synthetic data and tested on real data from our robot. We have observed that our approach does indeed show that it suitable for handling the robot-specific challenges of image segmentation, including camera shake and changing white balance.

For comparison to the rest of the segmentation field, trained our model on the PASCAL VOC 2012 image segmentation test. Our the qualitative results from our training progress over 16000 iterations over an ImageNet pre-trained model is shown in Figure 2 [4] [8]. We found our model ranked with the top models; our pixel accuracy was 91.71%, our mean accuracy was 82.29%, and our mean Intersection over Union (IoU), or Jaccard index, was 73.94%.

\[
J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \tag{1}
\]

The Jaccard index, shown in Equation 1, is a useful metric because it rewards pixel accuracy, but also penalizes pixel false positives. For comparison, the standard DeepLab segmentation algorithm off which our architecture is based gets 66% IoU, and the current state-of-the-art is 86% IoU on the PASCAL VOC leaderboard.

One challenge that we have observed at the moment is that, while our preliminary results are encouraging, the model is slow to train and evaluate, which could be potentially prohibitive implementation on robot operating in real time. We hypothesize that we can reduce this evaluation time constraint by shrinking the Region of Interest of the detector and using the dense CRF only as-needed.

V. CONCLUSIONS AND FUTURE WORK

In this work, we used synthetic data to improve the performance of a state-of-the-art deep neural net for semantic scene parsing, and integrated it into a pipeline that is more suited for robot vision. We demonstrated that this pipeline was able to...
cope with the error common in robot vision. While we have made progress in satisfactorily demonstrating the capabilities of our model in segmentation task on real scenes, we plan to further show this through tests on more datasets. The next steps for this work is to demonstrate explicit improvement on the ADE20K test dataset, showing robust frame-to-frame correspondence on validation video data from the SceneNet RGB-D dataset, and show the model trained on the synthetic data successfully segmenting the real robot camera data. We also are testing this network on the Cityscapes dataset as it has applications in autonomous driving. We have reconfigured the architecture to a two-stream network, where one network learns the optical flow, so that we can more explicitly take advantage of temporal information of the video data. We believe that video being sequences of frames and scenes being continuous in nature, the frames are highly correlated and can be used to aid the process of segmentation. Lastly, we plan to conduct an ablation study with different real training data supplemented by the SceneNet RGB-D synthetic data to show the degree of improvement from adding extra synthetic data to the training process.

By showing good results from this approach, we hope to motivate robots systems that are able to train and retrain their own vision systems by actively simulated scenarios and sampling them for new training examples, even while the robots are running.

ACKNOWLEDGMENTS

We would like to acknowledge the help and guidance given by Professor James Hays in this project, without whom this would not have been possible. We would also like to acknowledge the help of Shaurye Aggarwal annotating the robot data. Lastly, we would also like to thank Zhaoyang Lv and the members of the Wall Lab at Georgia Tech for their suggestions and help.

REFERENCES

[16] Pushmeet Kohli Nathan Silberman, Derek Hoiem and


