FeCCM for Scene Understanding: Helping the Robot to Learn Multiple Tasks

Congcong Li, TP Wong, Norris Xu, Ashutosh Saxena

***Abstract*— Helping a robot to understand a scene can include many sub-tasks, such as scene categorization, object detection, geometric labeling, etc. Each sub-task is notoriously hard, and state-of-art classifiers exist for many sub-tasks. It is desirable to have an algorithm that can capture such correlation without requiring to make any changes to the inner workings of any classifier, and therefore make the perception for a robot better. We have recently proposed a generic model (Feedback Enabled Cascaded Classification Model) that enables us to easily take state-of-art classifiers as black-boxes and improve performance.**

In this video, we show that we can use our FeCCM model to quickly combine existing classifiers for various sub-tasks, and build a shoe finder robot in a day. The video shows our robot using FeCCM to find a shoe on request.

I. INTRODUCTION

Consider building a simple robot that can fetch items on request, such as in Figure 1 or in some other prior works such as [12], [4]. This requires building several components—path planning, control, and so on. While most of these algorithms have been developed and easily available for use in open-source repositories (e.g., ROS [5]), perception still remains one of the big challenges—it is quite hard to build a reliable shoe detector on a short notice.

In particular, some state-of-the-art classifiers already exist for tasks such as scene categorization [11], scene geometric layout [8], object detection [3], and such. While the accuracy of a particular object detector may not be high enough, several other attributes can bring its accuracy to acceptable levels. For example, in Fig. 1, if we know that the scene is an office and we also know the general surface layout of the scene (i.e., where is the ground, walls, etc.), detecting a shoe becomes easier—because we know where to look for. Hence, if we are able to figure out a method to combine these state-of-the-art classifiers for different sub-tasks easily, robots have a hope to be able to perceive the environment.

Unfortunately, it turns out that previous works on combining classifiers are often daunting in that those methods tend to be adhoc and require a researcher to understand the inner workings of each classifier. For a roboticist, this could be even more difficult—not only one has to worry about getting algorithms such as planning and control working, but now he has to understand several different perception algorithms in order to build a robot just to fetch a shoe!

In our work, we have recently proposed generic learning algorithms called FeCCMs ([7], [10], [9]) using which one can quickly combine different classifiers together, without needing to know how each of them works. Our method considers each different perception algorithm as a "blackbox". That is, as long as each algorithm can produce an output given some inputs, and provide an interface to for its own training, our method can combine these sub-tasks in order to build a reliable classifier for a particular task.

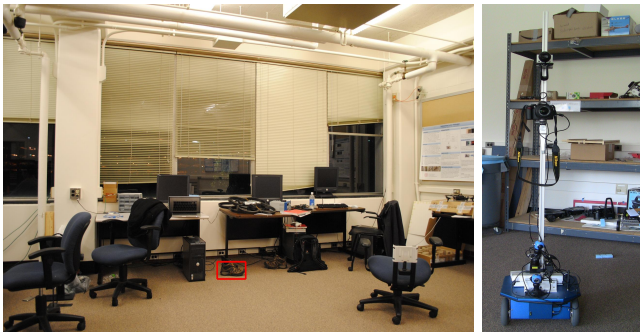

Fig. 1. Left: A mission for the robot: finding a shoe. Right: our robot. The robot takes advantage of scene categorization and geometric layout to aim shoe detection with our FeCCM algorithm [9].

In order to demonstrate the power of our approach, we took the challenging task of building a robot that can find a shoe on request. And we wanted to be able to do so within a day of computational time and within a day of human hours. We found the controllers for the robot and laser-based obstacle avoidance in ROS, but did not find a reliable shoe-finder code anywhere. One naive option would be to take the usual approach of collecting hundreds of thousands of shoe images, and then train a classifier, or use humans as domain experts to write hand-written rules about how to find a shoe better in an image. Instead, we just plugged in a standard off-the-shelf code for scene categorization, off-the-shelf code for geometric scene layout and off-the-shelf code for object detection. Even given only a few images to train with, we were quickly able to build a reliable shoe classifier.

This video submission does not allow us to present more quantitative results, but please refer to our algorithms and results described in [9] where we show that on six different vision tasks, our final outputs always improve performance.

We conduct experiments both offline and on robots. In the offline experiments, we show that we achieve improvements on each of these sub-tasks in indoor scene understanding: scene categorization, geometric layout, and object detection. Based on the learned model, we conduct experiments on our robot to complete a mission of finding a shoe in an indoor scene (as shown in Figure 1 and in the video).

Congcong Li is with Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA cl758@cornell.edu

TP Wong, Norris Xu and Ashutosh Saxena are with Department of Computer Science, Cornell University, Ithaca, NY 14853, USA {tw227, nx26}@cornell.edu, asaxena@cs.cornell.edu

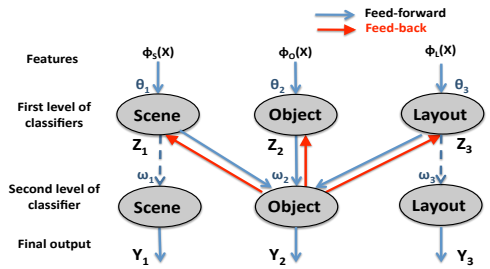

Fig. 2. FeCCM model for composing three sub-modules to aid indoor scene understanding. (Each task in the first layer is parametrized by θ_i and its output is Z_i, while each in the second layer is parametrized by ω_i with its output Y_i)

II. ALGORITHM

Our algorithm is composed of a 2-layer cascade with a hidden intermediate layer, which automatically learns connections between layers for optimizing the target task on the second layer [10] . Instead of learning a separate set of parameters for a specific sub-task, We propose an algorithm that optimizes the joint likelihood of all sub-tasks in [9]. In our model, the outputs from classifiers on the first layer go as input into the classifiers in the second layer. Figure 2 shows an example of using the proposed FeCCM model to compose scene categorization, object detection and geometric layout.

This learning algorithm can be trained with heterogenous datasets and with various perception algorithms. We treat each classier as a "black-box", with no restrictions on its operation other than requiring the ability to train on data and have input/output interface. This is often the case in robotics, where we want to quickly combine existing classifiers for building a robot for a particular task. We refer the reader to [10], [9] for more details on the algorithm, but we summarize how to use the algorithm for a particular robotic task:

1) **Step 1:** Download the code and datasets (whatever limited amount you can get) for a number of perception task related to your perception task.
2) **Step 2:** Train each classifier individually and estimate the parameters, using the code provided by them.
3) **Step 3.** Use the feedback step in [9] to estimate the latent variables.
4) **Step 4:** Go to step 2 and repeat for a few iterations.

Following these simple steps, we have shown that for a wide variety of tasks, we get significant improvements.

Robotic Tasks: This model can be widely used for robots that want to learn multiple tasks. Robots with different functions may need to combine a different group of classifiers. Developing a specific model to connect the tasks would need considerable insight into the inner workings of each task and also into how to connect/model each of them. However, with the generic FeCCM model, we can simply treat any existing state-of-the-art classifiers as black-boxes.

III. EXPERIMENTS

A. Offline Experiments

We combine three tasks for indoor scene understanding (in Figure 2): scene categorization, geometric layout and object detection. For scene categorization, we use the indoor scene subsets in the Cal-Scene Dataset [2] and classify an image into one of the four categories: bedroom, living room, kitchen and office. For geometric layout, we use the Indoor Layout Data [6] and assigning each pixel to one of three geometry classes: ground, wall and ceiling. For object detection, we use the PASCAL 2007 Dataset [1] and our own shoe dataset to learn detectors for four object categories: shoe, dining table, tvmonitor, and sofa. For the classifiers in the first layer, we use the scene classification algorithm in [11], the pixel-based geometry labeling algorithm in [8], and the part-based object detection algorithm in [3] for the corresponding tasks. Table I gives a summary of results for all three tasks. Our FeCCM method outperforms the base models (i.e. state-of-art methods) as well as the original CCM for most tasks.

TABLE I
SUMMARY OF RESULTS FOR THE THREE TASKS.

Model	Scene Categorization (% Accuracy)	Geometric Labeling (% Accuracy)	Object detection (% Average precision)			
			tv	table	sofa	shoe
Base-model	65.1	75.3	38.4	22.8	24.5	41.5
CCM [7]	65.1	84.9	38.9	22.6	24.4	42.1
FE-CCM	**66.7**	**86.3**	**39.1**	**23.0**	**24.5**	**43.8**

B. Robotic Experiments

we use the FeCCM model to build a shoe-finding robot (as shown in Figure 1 and in the video). With only a few training images, it is hard to train a robust shoe detector to find a shoe far away from the camera. However, the robot learns to take advantage of the other tasks through the FeCCM model and performs a more robust shoe detection.

In the video, we show that we can use our FeCCM model to quickly combine existing classiers for various sub-tasks, and build a shoe finding robot in a day. The video shows our robot using the trained FeCCM to find a shoe on request. Upon the user's request, the robot starts taking an image, and performs the tasks in the FeCCM model. It keeps moving and repeating this process until a shoe is detected. Based on detected position on the image, the robot calculates the shoe's physical position and navigates towards the shoe.

REFERENCES

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[2] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. *CVPR*, 2005.

[3] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2009.

[4] W. Garage. Beer Me, Robot. http://www.willowgarage.com/blog/2010/07/06/beer-me-robot, 2010.

[5] W. Garage and S. University. Robot operating system (ros). http://www.willowgarage.com/pages/software/ros-platform, 2008.

[6] V. Hedau, D. Hoiem, and D. Forsyth. Recovering the spatial layout of cluttered rooms. In *ICCV*, 2009.

[7] G. Heitz, S. Gould, A. Saxena, and D. Koller. Cascaded classification models: Combining models for holistic scene understanding. In *NIPS*, 2008.

[8] D. Hoiem, A. A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1), October 2007.

[9] C. Li, A. Kowdle, A. Saxena, and T. Chen. Feedback enabled cascaded classication models for scene understanding. In *NIPS*, 2010.

[10] C. Li, A. Kowdle, A. Saxena, and T. Chen. A generic model to compose vision modules for holistic scene understanding. In *ECCV Workshop on Parts and Attributes*, 2010.

[11] A. Oliva and A. Torralba. Mit outdoor scene dataset. http://people.csail.mit.edu/torralba/code/spatialenvelope/.

[12] A. Saxena, L. Wong, M. Quigley, and A. Y. Ng. A vision-based system for grasping novel objects. In *ISRR*, 2007.