Can you pick that? Problems with Vision for Robotic Pick-and-Place Applications

Abhijit Majumdar
Plus One Robotics
San Antonio, Texas 78226
abhijit.majumdar@plusonerobotics.com

Halit Bener Suay
Plus One Robotics
Pittsburgh, Pennsylvania 15213
bener.suay@plusonerobotics.com

Daniel H. Grollman
Plus One Robotics
Boulder, Colorado 80301
dan.grollman@plusonerobotics.com

I. Introduction

Robotic manipulators have been leveraged for industrial automation for several decades, mostly for tasks characterized by stable, predictable environments. More recently, improved sensing and control enabled these systems to operate in more dynamic environments, namely warehouse logistics. This domain is particularly challenging due to the stochastic nature of the environment and the high variability of the items (packaging) that need to be handled.

We present issues in identification and singulation of products in warehouse logistics and focus on some gaps between academic work in this field and the problems we encounter during product deployment. We list out the areas of needed focus and potential solutions for continued work to address these gaps.

II. Industry vs Research

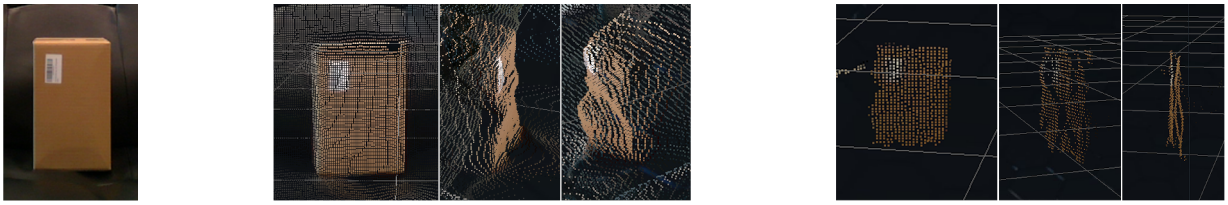
Fig. 1: Special packaging for what is typically shipped as a simple box, showing out-of-distribution data

Our domain of interest is automatic robotic manipulation of packages for pick-and-place tasks in warehouse logistics. We focus on a system with RGBD sensors that generate point clouds and color images. From this data we estimate the location and various attributes of individual packages to guide the manipulator for task execution. Attribute extraction is typically performed via deep learning methods, geometric / model-based computer vision methods, heuristic algorithms or a combination of all. While research has provided many techniques for approaching this problem, successful deployment into a product often encounters additional issues, described below:

- Sensor noise: Every sensor has noise, which can be dependent upon a number of environmental factors, such as distance to target, material, ambient light or temperature, etc. Further, for cameras, there is often a trade off between capture speed and noise - At the speeds required for production deployment, noise profiles are generally worse. Our observation is that such noise levels are more pertinent in the depth sensor than the color image sensors[1]. Further, each class of sensor often has its own noise profile, see Figure 2 for a comparison of two popular depth cameras. For repeatability, many research endeavors use simulated data, with either no noise, or a stationary noise model [2] When we use these techniques on our real data, they often fail. Most other research [4] that do present results on real world point cloud data use datasets from the autonomous driving domain that are not aligned with the logistics industry since the latter demands a more dense and refined definition of object features to be able to distinguish between multiple instances.
- Data distribution: Deep learning as a subset of machine learning is data driven, and hence it is often assumed that the test set for evaluating a model is drawn from the same distribution as the training set was for valid analysis of model performance. This is observed in most research on new deep learning model architectures [5]. However this dependence has impeding consequences when being used in the industry because of two factors:
  - Lack of large amount of labeled data for applications. This often causes data imbalance in training and test sets, since the most commonly used method of randomly splitting a dataset has a much lower probability of having similar testing instances in the training set.
  - Industry expectation of being able to handle any package. Collecting data is limited to the package types observed during data collection, which is very expensive when performed during production hours of the facility, therefore it is unrealistic to capture all package types for model training purposes. Special promotional packaging is a typical example of out-of-distribution packages introduced by the industry, see Figure 1.

  Emphasis should also be given to what data is considered to be within the same distribution. The typical distinction of a data point to be within distribution is indicated by the object of interest belonging to the same category [5].

[1] Depth sensing, being a newer technology, often has a greater noise problem than RGB sensing at similar price points



(a) Example package (b) D415 pointclouds of front and sides (c) L515 pointclouds of front and sides

Fig. 2: Different noise profiles from the Realsense D415 vs L515 sensors

However, deep neural networks do not mimic humans when encoding features from images, therefore seemingly negligible changes in images, can have unintended negative effects on the outcome of neural network predictions, and so a more feature-rich abstraction should be considered. Consider an example where a network is designed to detect 3 different types of packages one of which is a box, with training set containing images with brown boxes while the test set has brown boxes with tape on them. In this case the human distinction of boxes is negligible and would categorize any box within the same category, however for a neural network the tape plays a major role since such features were never observed in the training set. Hence, this should be considered a separate virtual category while splitting train and test sets.

- Deep learning model training cost: Packaging changes and long-tail problem for rare packages require improving the deep learning model frequently. As the dataset grows, the model performance improves, however the resources required to train the model also increases. Deep learning research often give less importance to the labeling and training resources required [1, 3] since their model iterations are only limited to obtaining one good model. However utilizing such training resources for model updates in a product is not a sustainable nor scalable solution for industrial applications.

III. Closing the Gap

To enable more efficient commercialization of research, we recommend the following:

- More research on depth sensors and their noise profiles, which can be used to improve signal-to-noise-ratio (SNR) for the sensors and detecting transparent materials like bags and glass. We also encourage researchers to publish the results for their deep learning models on real sensor data under various environmental conditions in addition to simulation data for developers to get a better understanding on fit for a method into the application.
- We believe that it should be made a practice for deep learning research to always include performance metrics for out-of-distribution data. This not only exposes caveats in the model generalization capabilities but also indicates to a developer on the potential of use of such model in their application as it would address the industry demand for system to perform at-par for new packages up to a certain threshold outside the training data distribution.

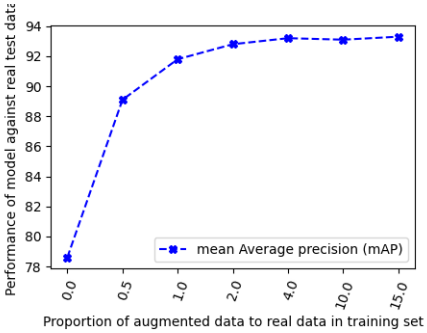


Fig. 3: Effect of different amount of augmentation on model performance. Improves performance until a saturation point

- When inferring on a variety of product types, observing data distribution becomes key when long-tail problems arise. Model evaluations and solutions to such problem are important when being used in production because these are typically the most difficult to debug and collect data for. Researchers are encouraged to consider model performance evaluations under unevenly distributed data and potentially derive solutions to such cases.
- Augmentation was found to be a major key in solving for unprecedented scenarios, to account for the lack of substantially large datasets and hence the variability in collected data. This largely affects generalization of the model which can be targeted towards the application at hand. Figure 3 shows our experiments run on limited data (1000 real images) and how augmentation (copy-paste, augmix, geometric transformations, etc.) improves performance. Additionally we would like to encourage research on simulating camera imaging capabilities to augment data from specific sensors used in production not idealistic images.
- We believe research should not only be focused on making inferences faster [6, 7] but also on the ability to train deep neural networks faster and more efficiently with less resources.

IV. Conclusion

We have presented observations from putting cutting-edge deep-learning based vision systems into practice for pick-and-place manipulation in warehouse logistics. While the research community continues to advance the state-of-the-art, we have found several stumbling blocks that bear further investigation. It is our hope that renewed interest in these areas will prove advantageous to both communities.

REFERENCES

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020.

[2] B. Deng, K. Genova, S. Yazdani, S. Bouaziz, G. E. Hinton, and A. Tagliasacchi. Cvxnets: Learnable convex decomposition. *CoRR*, abs/1909.05736, 2019.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[5] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

[6] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016.

[7] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. *CoRR*, abs/1911.09070, 2019.