Representing Visual Scenes for Robot Control

Dr Zeeshan Zia

May 25, 2015

Abstract

The objective of the proposed line of research is to develop new knowledge representation frameworks that can model a variety of visual scenes at multiple levels of detail - and leverage them to significantly advance the state-of-the-art in robot control. On one hand, computer vision has recently seen many advances from dense SLAM approaches to object class recognition and semantic scene modeling. These advances mean that vision based techniques can no longer be neglected by the robotics community. On the other hand, results from deep neural network research have shown that being able to leverage large amounts of training data also holds promise for mid-level control problems such as navigation and manipulation, as well as greatly enhancing the applicability of reinforcement learning. My aim is to invent and leverage rich 3D scene representations tightly integrated with high fidelity physics simulation techniques to generate training data for deep recurrent neural networks, and in turn utilize the learned models for robust and versatile manipulation and navigation.

1 Introduction and Open Questions

I am fascinated by recent achievements in visual mapping and recognition and see clear opportunities in combining these with data-driven methods for robotic control. While the title of the proposal, tackling manipulation as well as navigation together might seem too ambitious, recent advances in deep neural networks have shown that very similar techniques can be applied to solve a variety of problems across widely separated domains. Manually engineered mid-level robotic control systems have been researched thoroughly over the past two to three decades, without much success for out-in-the-wild robotics applications, such as indoor service robots or driverless vehicles operating in the absence of heavily annotated maps. Recently successful machine learning approaches which can learn extremely complicated relationships between various high-dimensional input and output variables have the potential to alleviate these problems, opening the possibility for highly versatile and robust mid-level robotic control. I feel uniquely suited to perform (and supervise) this research given my experience in both machine learning applied to computer vision and software-hardware auto-tuning and high-performance computation. In the following, I highlight some opportunities and open questions in robot vision and control systems.

1.1 Robot Vision

A number of representations emphasizing different levels of detail have been proposed for scenes [18, 27, 37, 8, 29, 19], both from the semantic and the geometric vision communities. These representations emphasize either the semantic or the geometric aspects of the scene, e.g. a room cuboid with surface aligned bounding boxes to represent some of the obj ects or a Truncated Signed Distance Function (TSDF) voxel grid to represent surfaces. A superior scene representation would be one which could combine both these aspects into one fluid model - fluid because in real scenes objects can be moving and often the concept of object itself is not well-defined. I take inspiration from a video game scene representation called the Scene graph [4]. *How can we compute the "inverse scene graph" given a stream of 2.5D images? How to define an optimal trajectory to refine the representation (active vision)?*

Visual object detection has a rich history [7, 30, 23, 21, 34, 16, 32, 35, 12, 10, 33, 14], and is the most widely researched sub-problem in computer vision. Recently, large advances in detection performances have been made possible by the revival of deep convolutional neural networks [33, 14] relying on very large labeled datasets and exponential growth in computational resources (consumer GP-GPU devices). However, still the results remain far behind practically useful levels, at around 40% mean AP atleast in unconstrained

settings. One of the fundamental nuisances for computer vision - the loss of information caused due to projecting a 3D scene onto a 2D plane - and its consequences specially, ambiguity in segmentation, mean that there is a lot of potential for exploiting RGB-D images in conjunction with deep CNNs. Unfortunately, the large amount of (labeled) data - multiview depth images labeled at the level of object identities, geometries, and viewpoints - needed for such an enterprise do not exist. *How can we exploit 3D reconstructions coupled with graphics rendering techniques and enriched with 3D CAD data to generate large quantities of finely labeled training (RGB+Depth) data?*

On the side of geometric computer vision, dense SLAM [28, 27, 37, 8, 29, 19] is the problem of building rich 3D descriptions (as opposed to sparse keypoint-based representations) of the environment while estimating the location of the sensor at each point in time. Some of the most recent approaches in dense SLAM allow reconstructing detailed models of scenes, which can then be rendered from any viewpoint, and allow for segmentation exploiting both photometry and 3D geometry. These reconstructions can be useful for an indoor service robot or for an augmented reality pipeline. Unfortunately, these methods do not give any semantic meaning to the geometric structures, which significantly limits their applicability. *How can we simultaneously perform object detection together with the 3D reconstruction? What are the representations at the level of data structures and computations that will make these methods scalable in terms of numbers and types of objects?*

1.2 Manipulation and Navigation

A lot of work has been done on mid-level robotic control problems such as manipulation and navigation over the past two decades. Most of this work [9, 36, 31, 5] has focused on explicitly trying to perform state estimation and accordingly deriving control policies in either a deterministic or probabilistic framework. While this work has resulted in applications for relatively controlled settings; manipulation of different kinds of objects in an unconstrained setting, or locomotion in cluttered environments remain largely unsolved [2]. By and large, researchers have tried to focus on special cases of the manipulation and navigation problem and developed techniques to solve those special cases with varying degrees of success. However the question of how to optimally combine these interacting sub-systems into a full system in a robust way remains wide-open. Recently, the advances in GPGPU computation and the availability of large amounts of training data in some domains such as visual classification and speech recognition have enabled successes that were previously unheard of - while requiring minimal explicit modeling. Relatively few of these ideas have flowed into the robotics control domain mainly due to the unavailability of vast quantities of training data and because these are still early days in successful dissemination of dense SLAM and detailed semantic scene understanding results among the robotics control community.

The few early attempts include the LAGR system [17] for autonomous navigation which learns the mapping from raw input images to steering wheel control without requiring any explicit SLAM module. The system currently is not able to handle other moving vehicles, and works only for fixed (albeit fairly complex) scenes. This is because it will require many orders of magnitude more training data to learn the control with multiple other moving vehicles, directly from raw images. On the contrary, a more amenable higher level scene representation could reduce the need for training data significantly. Along similar lines,[22] propose a pipeline for robotic manipulation by learning direct mapping from image data to motor control signals using convolutional neural networks. Unfortunately the learned policies are specific to one kind of scene and object here, and not transferable to other kinds of scenes. For example, the system fails if the scene background changes or if the object dimensions change. Again a more sophisticated representation of the scene and 3D object geometries (as opposed to 2D images) fed as input to the deep learning pipeline could allow far more versatile capabilities. Another widely acknowledged work is about deep reinforcement learning for playing atari games [25], enabled by tons of training data generated through simulation. While a major weakness of the system here includes its limited memory and reasoning ability, the deep learning agent could nonetheless already be unleashed in a more realistic 3D environment to see how it would cope. One route would be still to do this learning in the context of a 3D game; however I believe a good physics simulator coupled with mixed reality graphics enabled by dense SLAM approaches and 3D CAD data could more directly enable realistic application. *How can physical simulators be efficiently coupled with mixed*

scene synthesis pipelines? How to make learning in such settings computationally feasible with available GPGPU hardware?

While ideally the learning algorithm should be able to learn the appropriate representation suited to the task from training data; in practice still a lot of hand tuning needs to be performed to guide that learner towards the right kind of features. There is already evidence [6] that providing richer scene representations, here simply in the form of optical flow, enhances prediction accuracy. Thus I stongly believe that representing the visual input as explicit rich higher-level representations comprising of geometry and semantics can significantly improve deep reinforcement learning pipelines. *How can a detailed hierarchical 3D representation of the scene be coupled with a deep neural network?*

2 Relation to my background

My research so far has focused on holistic scene understanding [38, 44, 41, 40, 42, 39, 24, 43, 26]: working on both indoor and outdoor scenes, for static and dynamic settings, both for computational/power constrained and unconstrained setups. I have thought a lot about where real-time 3D scene understanding is going, and tried to pin down the resources that will be available in the future by mobilizing industry and academic leaders in the domain [1]. Apart from my core expertise in robot vision, I also had the opportunity of working with researchers who have contributed a lot to robot manipulation and planning (group of Prof. Michael Beetz at TU Munich, CoTeSys research cluster). This experience gives me contextual understanding of the problems and successful approaches in these domains. I also possess an indepth understanding of various machine learning approaches as they have been applied to both computer vision and mid-level robot control.

This experience makes me well-suited to conduct and supervise the research proposed herein.

3 Societal Impact

Robust visual scene understanding and versatile control capabilities have the potential to completely change the way we live for the better. Applications enabled and made practical will include:

- Domestic service robots: from manipulation to path planning, enabling assistive households for the elderly.
- Industrial robotics: warehouse automation, as an example of robust sensing in uncontrolled environments.
- Self-driving cars: enable navigation without the need for high-resolution manually annotated reconstructions.
- Augmented Reality: enable new user interfaces and use-cases for ubiquitous computing.

4 Research Project

Concretely, I plan on initiating work on the following projects. Together with myself, I plan on supervising graduate students and postdoctoral researchers funded by European and Norwegian research grants. I know that NTNU already possesses a strong robotics group, and I would be looking forward to collaborating with them specially on the control aspect of these projects, and hopefully also on existing projects.

4.1 Hierarchical Scene Representations

While it is common to talk about objects in the context of visual recognition, the concept is not well-defined [11]. Is my desk an object? or are its legs the objects? or is my desk together with my computer on it one object? A natural representation for the 3D scene would comprise of these “object” hierarchies -

including provisions to modify the representation in case something moves or as we get more observations (in the case of a moving camera exploring an environment).

Objects are defined in [11] on the basis of grouping processes; objects are what pop out as the "ultimate" products of such processes. I imagine a tree representation of a set of images - which might be algorithmically implemented as divisive clustering, using deep learning based segmentation on the RGB channel coupled with 3D proximity.

We would start by recursively sub-dividing the scene: individual RGB-D images in a point-based fusion style dense SLAM framework [19], into a hierarchy based on pixel/surfel grouping. The root node would comprise of the entire scene, and an objective function comprising of our grouping cues (based on current view of the scene) would be evaluated to recursively sub-divide the surfel cloud. The edges of the tree would represent rigid transformations between these nodes. Planar segments, which are important components for most man-made scenes, would automatically pop out as a result of such segmentation. It is expected that in such a representation, my desk's leg would automatically be a sub-tree, and if I cut the sub-tree at one higher node, I will get my desk, another branch up and I might get my desk together with all the clutter on top of it. A related idea in video games is called the Scene Graph [4] - we will be computing the "inverse scene graph".

Doing bundle adjustment in the SLAM context on such a representation should be very cheap - since the rigid transform has to be computed and maintained for the 3D segments rather than individual (thousands of) surfels. Also having such an explicit hierarchical representation would allow us to perform learning and recognition - we could modify the representation in ways such as compress a sub-tree into a single leaf node based on prior knowledge etc. Further, since a single tree would probably never be perfect, why not have a forest of scene representations, which can be thought as multiple hypotheses for the scene. Each tree could yield a slightly different hierarchy which captures the inherent uncertainty/ambiguity about inter object relations/dependencies. Further, such a representation could allow functional relationships such as "is a" and "has a" to be readily captured.

4.2 RGB + Depth CNNs, Detailed Object Representations, Dense SLAM

Deep CNNs have been shown to improve object detection performance [33, 14] on RGB images. However, relatively little research has been done on training CNNs on RGB + depth images with the aim of detailed scene understanding, due to lack of large *finely* labeled training sets. Coarse grained 2D bounding box level detections are investigated in [15], but 2.5D images provide the opportunity for more detailed analysis of the detected object, e.g. direct estimation of object viewpoint and fine-grained classification. I believe that Dense SLAM approaches combined with graphics rendering techniques provide an excellent opportunity to generate massive amounts of such training data, together with fine-grained poses and segmentation masks. It is possible to further increase the size of such sets by synthetically varying lighting conditions and rendering 3D CAD data from CAD model databases. Weaker detectors and segmentation routines can be used to bootstrap a manual annotation process which can reduce the annotation effort by a large amount. CNNs learnt on such richly labeled RGB+D data would be far superior for object detection tasks.

4.3 Learning Mid-Level Robot Control

Detailed explicit scene representations incorporating geometry and semantics can allow superior deep neural networks to be learned for robotic control, as opposed to raw input images. We would start by utilizing pre-existing graphics pipelines [13] coupled with real world 3D reconstructions, and loaded into a state-of-the-art physics engine [20]. Incorporated into the simulated world would be a physical agent such as a PR2 robot [3]. The agent will physically try out a range of actions in a randomized sampling framework, all the while, recording the simulated data for training a deep reinforcement learning pipeline. The simulated scene would be encoded in terms of the hierarchical scene representation described above both for training and testing. Not only do I expect the vision to help control, but also the other way round. One demonstration scenario might be that of a PR2 robot operating inside an assistive kitchen: cleaning dishes, cooking meals, and moving around the kitchen in doing so.

5 Conclusion

I am interested in combining ideas from semantic and geometric computer vision for understanding complex dynamic indoor scenes. In particular, I am interested in bringing together evidence from both types of cues into a coherent and flexible representation. RGB-D cameras and deep neural networks have been shown useful for many applications in computer vision, however, they have not been exploited enough for holistic scene level understanding or active vision in real-time settings. I hope to alleviate this discrepancy through my research.

There is a lot of opportunity in robotic control to exploit recent advances in vision, machine learning, and GPGPU technology, which can enable versatile and robust application of robots for tasks such as manipulation and navigation. I am particularly interested in demonstrating the developed techniques in the context of an indoor service robot assisting humans in their daily routine.

References

[1] BMVA Workshop on Real-time 3D Scene Understanding in year 2020 (17th June 2015). http://www.zeeshanzia.com/y2020.htm.

[2] DARPA Robotics Challenge Finals 2015 (Accessed: 20 May 2015). http://www.theroboticschallenge.org.

[3] PR2 Robot Simulator plugin for ROS (Accessed: 20 May 2015). http://wiki.ros.org/pr2simulator/Tutorials.

[4] Scene graph (wikipedia article, 15 april 2015). http://en.wikipedia.org/wiki/Scene-graph.

[5] Michael Beetz, Ulrich Klank, Ingo Kresse, Alexis Maldonado, L Mosenlechner, Dejan Pangercic, T Ruhr, and Moritz Tenorth. Robotic roommates making pancakes. In *Humanoids*, 2011.

[6] R. Benenson, M. Omran, J. Hosang, , and B. Schiele. Ten years of pedestrian detection, what have we learned? In *ECCV, CVRSUAD workshop*, 2014.

[7] R. A. Brooks. Symbolic reasoning among 3-d models and 2-d images. *Artificial Intelligence*, 17(1-3):285–348, 1981.

[8] Jiawen Chen, Dennis Bautembach, and Shahram Izadi. Scalable real-time volumetric surface reconstruction. *ACM TOG*, 2013.

[9] Mark Cutkosky. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *IEEE Transactions on Robotics and Automation*, 1989.

[10] Mark Everingham, Luc Van Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[11] J Feldman. What is a visual object? *Trends in Cognitive Sciences*, 2003.

[12] Pedro F. Felzenszwalb, Ross Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[13] Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas Funkhouser, and Pat Hanrahan. Example-based synthesis of 3d object arrangements. In *SIGGRAPH Asia*, 2012.

[14] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[15] Saurabh Gupta, Ross Girshick, Pablo Arbelaez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. 2014.

[16] Michael Haag and Hans-Hellmut Nagel. Combination of edge element and optical flow estimates for 3d-model-based vehicle tracking in traffic image sequences. *IJCV*, 35(3):295–319, 1999.

[17] Raia Hadsell, Pierre Sermanet, Jan Ben, Ayse Erkan, Marco Scoffier, Koray Kavukcuoglu, Urs Muller, and Yann LeCun. Learning long-range vision for autonomous off-road driving. *Journal of Field Robotics*, 2009.

[18] D. Hoiem and S. Savarese. *Representations and Techniques for 3D Object Recognition and Scene Interpretation*. Morgan and Claypool Publishers, California, USA, 2011.

[19] Maik Keller, Damien Lefloch, Martin Lambers, Shahram Izadi, Tim Weyrich, and Andreas Kolb. Real-time 3D reconstruction in dynamic scenes using point-based fusion. In *3DTV-Conference*. IEEE, 2013.

[20] Nate Koenig and Andrew Howard. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *IROS*, 2004.

[21] D. Koller, K. Daniilidis, and H. H. Nagel. Model-based object tracking in monocular image sequences of road traffic scenes. *IJCV*, 10(3):257–281, 1993.

[22] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-End Training of Deep Visuomotor Policies. In *ICRA*, 2015.

[23] D. G. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 1987.

[24] Emilio Maggio, M. Zeeshan Zia, Qi Pan, Michael Gervautz, and Zsolt Szalavari. Component-Based Target Object Detection and Color Classification. US Patent Application, 2014.

[25] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing Atari with Deep Reinforcement Learning. In *NIPS Deep Learning Workshop*, 2013.

[26] Luigi Nardi, Bruno Bodin, M. Zeeshan Zia, John Mawer, Andy Nisbet, Paul H. J. Kelly, Andrew J. Davison, Mikel Luján, Michael F. P. O'Boyle, Graham Riley, Nigel Topham, and Steve Furber. Introducing SLAMBench, a performance and accuracy benchmarking methodology for SLAM. In *ICRA*, May 2015.

[27] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *ISMAR*, 2011.

[28] Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. DTAM: Dense tracking and mapping in real-time. In *IEEE International Conference on Computer Vision*, 2011.

[29] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM TOG*, 2013.

[30] A. P. Pentland. Perceptual organization and the representation of natural form. *Artificial Intelligence*, 28(3):293–331, 1986.

[31] N Ratliff, M Zucker, JA Bagnell, and S Srinivasa. CHOMP: Gradient optimization techniques for efficient motion planning. In *ICRA*, 2009.

[32] Henry Schneiderman and Takeo Kanade. A statistical method for 3d object detection applied to faces and cars. In *CVPR*, 2000.

[33] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *CoRR*, abs/1312.6229, 2013.

[34] G. D. Sullivan, A. D. Worrall, and J.M. Ferryman. Visual object recognition using deformable models of vehicles. In *IEEE Workshop on Context-Based Vision*, 1995.

[35] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. Van Gool. Towards multi-view object class detection. In *CVPR*, 2006.

[36] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.

[37] T. Whelan, M. Kaess, M.F. Fallon, H. Johannsson, J.J. Leonard, and J.B. McDonald. Kintinuous: Spatially Extended KinectFusion. In *WS on RGB-D: Advanced Reasoning with Depth Cameras*, 2012.

[38] M. Zeeshan Zia. Inside-Out Activity Analysis using 3D Hand, Object, and Scene Tracking. Master's thesis, TU Munich, 2009.

[39] M. Zeeshan Zia, Emilio Maggio, Qi Pan, Michael Gervautz, and Zsolt Szalavari. Exemplars-Based Color Classification. US Patent Application, 2014.

[40] M. Zeeshan Zia, Michael Stark, Bernt Schiele, and Konrad Schindler. Detailed 3d representations for object recognition and modeling. *PAMI*, 2013.

[41] M. Zeeshan Zia, Michael Stark, and Konrad Schindler. Explicit occlusion modeling for 3D object class representations. In *CVPR*, 2013.

[42] M. Zeeshan Zia, Michael Stark, and Konrad Schindler. Are Cars Just 3D Boxes? Jointly estimating the 3D Shape of Multiple Objects. *CVPR*, 2014.

[43] M. Zeeshan Zia, Michael Stark, and Konrad Schindler. Towards scene understanding with detailed 3d object representations. *IJCV*, 2015.

[44] M. Zeeshan Zia, Michael Stark, Konrad Schindler, and Bernt Schiele. Revisiting 3D geometric models for accurate object shape and pose. In *ICCV-WS 3dRR*, 2011.