Multitask Learning for Reliable State Estimation

Noha Radwan and Wolfram Burgard
Department of Computer Science, University of Freiburg, Germany

I. Introduction

An ultimate goal of mobile robotics research is the ubiquitous deployment of intelligent platforms that are capable of undertaking a variety of tasks in everyday life for their users. Over the previous decade, robots have become more integrated into our daily lives, performing tasks in numerous environments including industrial settings such as in assembly and manufacturing, indoor scenarios such as home assistance and elderly care, as well as outdoor tasks such as lawn mowing and parcel delivery. Despite the significant strides achieved in the various application areas, reliably deploying robots in urban environments remains an open challenge due to the complex and highly dynamic nature of the environment that renders hand-crafted solutions infeasible. In order to attain the goal of ubiquitous robotic deployment, robots need to accurately estimate their position as well as the position of other pedestrians or agents in their vicinity in order to ensure safe operation. In my work, I have focused on addressing the challenging problem of reliable and accurate state estimation in urban environments by introducing techniques that leverage the abundantly rich semantic, structural and geometric information in the scene. In the following sections, I briefly describe some of the challenges, proposed solutions and future research directions in the domains of localization and motion prediction.

II. Multitask Learning for Visual Localization

Visual localization is an essential enabler for various robotics and computer vision tasks such as Simultaneous Localization and Mapping [21], Augmented Reality [11], and autonomous navigation [6]. In order for robots to be safely deployed in the wild, their localization system should be robust to frequent changes in the environment, including seasonal changes, dynamic changes such as moving vehicles, and structural changes such as constructions.

While local feature-based approaches that utilize SfM information for localization [19, 26] achieve for the most part state-of-the-art performance, failures often occur with large viewpoint changes and motion blur. On the other hand, although deep learning-based methods are able to handle challenging perceptual conditions, they are still unable to match the performance of state-of-the-art local feature-based localization methods. This is partly due to their inability to model the 3D structural constraints of the environment while learning from a single monocular image. To address this shortcoming, we proposed a novel loss function that enables embedding the geometric knowledge of the scene by leveraging auxiliary learning to jointly estimate the ego-motion of the robot [25]. Our proposed loss function augments the Euclidean loss by the inclusion of an additional term that constrains the predicted poses to be consistent with the ego-motion of the robot. This improves the accuracy of the poses predicted by our network, as well as the robustness to perceptual changes (see Fig. 1).

TABLE I
Comparison on the Microsoft 7-Scenes benchmark.

Method	Median Error	Pose Acc.	Run-time
DSAC2 [4]	0.04m, 1.04°	76.1%	200ms
Ours	**0.013**m, **0.77**°	**99.2**%	**79**ms

The ability of the localization system to identify the stable features in the environment is key to enabling successful and reliable localization. Towards this goal, we propose to simultaneously predict the semantics of the surroundings as a means to instill structural cues about the environment into the pose regression network and implicitly draw more attention towards informative regions in the scene [15]. In order to facilitate the effective combination of feature maps across the task-specific networks, we proposed a novel layer that utilizes feature map activations to dynamically weigh the different semantic representations. We employ this layer to fuse feature maps from the segmentation stream into the localization stream and vice versa.

In order to enable the prediction of consistent semantics, we further propose a novel self-supervised semantic context aggregation technique that leverages the predicted relative motion from the odometry stream of our network [15]. Using pixel-wise depth predictions from a CNN and differential warping, we fuse intermediate network representations from the previous timestep into the current frame using our proposed fusion layer. This enables our segmentation network to aggregate more scene-level context, thereby improving the performance and leading to faster convergence. Through incorporating the semantic and ego-motion information into the localization task, our network architecture learns to jointly estimate all three tasks in a multitask learning (MTL) manner. Given two consecutive monocular images, our network predicts the 6-DoF global pose, ego-motion and semantic segmentation.

Exploiting MTL, our architecture achieves state-of-the-art performance on the challenging Microsoft 7-Scenes benchmark [20], while simultaneously preforming multiple tasks. Tab. I shows a quantitative comparison on the 7-Scenes dataset with the previous state-of-the-art method in terms of median localization accuracy, pose accuracy in terms of percentage of poses with error below 5cm and 5° and run-time. The results show that not only does our approach exceed DSAC2 [4] by 67.5% in the translational and 25.9% in the rotational components of the pose, it also improves on the pose accuracy by 23.1% and the run-time by 60.5%. This renders our method well suited for real-time deployment in an online manner.

III. Multimodal Motion Prediction

The ubiquitous deployment of mobile robots in urban environments necessitates the development of robust behavior

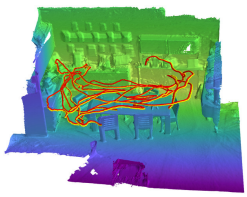

Fig. 1. Localization results depicting the predicted pose (yellow trajectory) versus the ground-truth pose (red trajectory) on the Red-Kitchen scene [20].

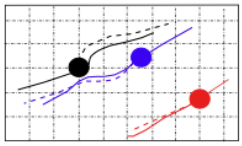

Fig. 2. Motion prediction results on the UCY-Uni dataset [10]. The solid line depicts the ground-truth trajectory and the dotted line depicts the network prediction.

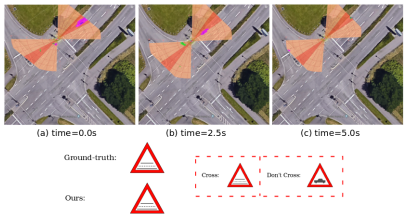

Fig. 3. Sequence images from an example street crossing scenario. Vehicles are represented by arrows where the size of the arrow is proprotional to the velocity of the vehicle.

TABLE II
BENCHMARKING THE MOTION PREDICTION ON UCY-UNI [10].

Method	Avg. Error (m)	Final Error (m)	Run-time (s)	Size (MB)
Social-LSTM [1]	0.27	0.77	1.78	95.8
SGAN [7]	0.60	1.26	0.04	N/A
Ours	0.29	**0.46**	0.06	**7.0**

prediction algorithms to ensure the safety of surrounding humans. Among the situations in which the behavior of the robot is crucial for the safety of itself and surrounding agents is navigating street intersections. The complexity of the scene and the presence of multiple dynamic objects render this problem extremely challenging [13].

Most existing robotic approaches either propose the use of vehicle-to-vehicle communication [17, 8] or rely solely on the traffic light signal [2]. Employing vehicle-to-vehicle communication methods requires the standardization of protocols among all manufacturers, which is infeasible. Similarly, solely relying on the traffic light information to make the crossing decision is suboptimal as not only is the traffic light recognition task challenging, the signal alone does not ensure the intersection safety for crossing.

In order to address this problem, in my previous work, we proposed a multimodal framework for jointly estimating the future trajectories of the observable traffic participants, recognize the state of the traffic light if present, and identify the safety of the street intersection for crossing [14]. Through leveraging the predicted trajectories of the observable vehicles and pedestrians in the vicinity of the robot, in addition to the state of the traffic light, our MTL framework is able to accurately estimate the safety of the street intersection for crossing, while being intersection invariant. Furthermore, as our approach does not rely on any prior knowledge of the environment or form of communication with the surrounding traffic participants, it can be easily deployed in various environments.

Unlike previous methods for behavior prediction [18, 12, 3], we proposed a novel scalable neural network architecture that employs causal convolutions to model the sequential behavior of the observable traffic participants. Our proposed architecture simultaneously predicts the trajectories of all observable traffic participants, thus enabling it to better leverage the interdependencies in their motion without the need for explicitly defining the relative importance between the various participants. Furthermore, we predict the heading (theta angle) of the observed dynamic objects, which enables our network to predict more accurate trajectories (see Fig. 2).

Concurrently, we proposed a convolutional neural network architecture for traffic light recognition that utilizes the global information in the images to selectively emphasize informative features and suppress irrelevant features using SE-blocks [9], thereby being more robust to noise in the input image. In order to learn a classifier that is robust to the type of intersection, we fuse the learned representations from the traffic light recognition network and the interaction-aware motion prediction network to infer the final crossing decision. By incorporating the uncertainty information from the motion prediction stream and the learned representations from the traffic light recognition stream, the classifier is robust to incorrect predictions by either sub-network. We evaluated our architecture on several indoor and outdoor datasets for motion prediction, traffic light recognition and street crossing prediction. Tab. II shows a comparison of the performance of our framework with state-of-the-art methods on the motion prediction task for the UCY-UNI dataset [10]. The results demonstrate that our approach improves upon the final displacement error by 40.3% while achieving analogous average displacement error with a competitive run-time of 0.06s. Moreover, our model requires only 7.0MB of storage space, thereby making it efficiently deployable in resource limited systems. Fig. 3 depicts an example crossing scenario where using our proposed approach, the network is able to accurately predict the safety of the interval for crossing as the oncoming vehicles slow down to a halt.

IV. FUTURE WORK

For future work, I plan to extend our work to address the problem of localization in dynamic environments. As the robot traverses in urban cities, it is often surrounded by other dynamic objects with different velocities than its own. Learning to estimate the motion of the surrounding dynamic objects in the scene through scene flow can be beneficial towards improving the ego-motion estimated by the network [16, 27]. Furthermore, the presence of dynamic objects in the scene can impair the quality of the predicted global poses. Learning to segment out the dynamic objects in the scene and inpaint the occluded parts of the image using GANs would enable our method to produce more accurate pose estimates [5].

In the context of motion prediction, I plan to incorporate the prediction of obstacle maps into the motion prediction sub-network [12]. Knowledge about the vicinity can substantially improve the accuracy of the predicted trajectories by avoiding paths that intersect with obstacles. Learning to semantically classify the traffic participants can also aid in understanding the potential interactions among them, thereby increasing the accuracy of the predicted trajectories.

Overall, I believe that developing frameworks that incorporate information from diverse tasks via MTL will improve the accuracy of the individual tasks by leveraging the similarities

and underlying interdependencies among the tasks and hence increase robustness to complex urban scenes.

REFERENCES

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] A. Bauer, K. Klasing, G. Lidoris, Q. Mühlbauer, F. Rohrmüller, S. Sosnowski, T. Xu, K. Kühnlenz, D. Wollherr, and M. Buss. The autonomous city explorer: Towards natural human-robot interaction in urban environments. *Int. Journal of Social Robotics*, 1(2):127–140, 2009.

[3] U. Baumann, C. Glaeser, M. Herman, and J. M. Zöllner. Predicting ego-vehicle paths from environmental observations with a deep neural network. In *Int. Conf. on Robotics & Automation (ICRA)*, 2018.

[4] E. Brachmann and C. Rother. Learning less is more-6d camera localization via 3d surface regression. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] H. Dhamo, K. Tateno, I. Laina, N. Navab, and F. Tombari. Peeking behind objects: Layered depth prediction from a single image. *arXiv preprint arXiv:1807.08776*, 2018.

[6] C. Gómez, M. Mattamala, T. Resink, and J. Ruiz del Solar. Visual slam-based localization and navigation for service robots: The pepper case. *arXiv preprint arXiv:1811.08414*, 2018.

[7] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] G. Habibi, N. Jaipuria, and J. P. How. Context-aware pedestrian motion prediction in urban intersections. *arXiv preprint arXiv:1806.09453*, 2018.

[9] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 2017.

[10] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *Computer Graphics Forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[11] E. Marchand, H. Uchiyama, and F. Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22 (12):2633–2651, 2016.

[12] M. Pfeiffer, G. Paolo, H. Sommer, J. Nieto, R. Siegwart, and C. Cadena. A data-driven model for interaction-aware pedestrian motion prediction in object cluttered environments. 2018.

[13] N. Radwan, W. Winterhalter, C. Dornhege, and W. Burgard. Why did the robot cross the road? - learning from multi-modal sensor data for autonomous road crossing. In *Int. Conf. on Intelligent Robots and Systems (IROS)*, 2017.

[14] N. Radwan, A. Valada, and W. Burgard. Multimodal interaction-aware motion prediction for autonomous street crossing. *arXiv preprint arXiv:1808.06887*, 2018.

[15] Noha Radwan, Abhinav Valada, and Wolfram Burgard. Vlocnet++: Deep multitask learning for semantic visual localization and odometry. *IEEE Robotics and Automation Letters (RA-L)*, 3(4):4407–4414, 2018.

[16] A. Ranjan, V. Jampani, K. Kim, D. Sun, J. Wulff, and M. J. Black. Adversarial collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. *arXiv preprint arXiv:1805.09806*, 2018.

[17] J. Rios-Torres and A. A. Malikopoulos. A survey on the coordination of connected and automated vehicles at intersections and merging at highway on-ramps. *Transactions on Intelligent Transportation Systems*, 2016.

[18] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, and S. Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. *arXiv preprint arXiv:1806.01482*, 2018.

[19] T. Sattler, B. Leibe, and L. Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *Transactions on Pattern Analysis & Machine Intelligence*, (9):1744–1756, 2017.

[20] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[21] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.

[22] A. Valada and W. Burgard. Deep spatiotemporal models for robust proprioceptive terrain classification. *Int. Journal of Robotics Research (IJRR)*, 36(13-14):1521–1539, 2017.

[23] A. Valada, G. Oliveira, T. Brox, and W. Burgard. Towards robust semantic segmentation using deep fusion. In *Robotics: Science and Systems (RSS 2016) Workshop, Are the Sceptics Right? Limits and Potentials of Deep Learning in Robotics*, 2016.

[24] A. Valada, R. Mohan, and W. Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *arXiv preprint arXiv:1808.03833*, 2018.

[25] A. Valada, N. Radwan, and W. Burgard. Deep auxiliary learning for visual localization and odometry. In *Int. Conf. on Robotics & Automation (ICRA)*, 2018.

[26] J. Valentin, M. Niener, J. Shotton, A. Fitzgibbon, S. Izadi, and P. Torr. Exploiting uncertainty in regression forests for accurate camera relocalization. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[27] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.