Dynamic Scene Understanding for Mobile Robot Navigation

Ondrej Miksik*
Department of Control and Instrumentation
Faculty of Electrical Engineering and Communication
Brno University of Technology
Kolejni 4, Brno, Czech Republic
ondra.miksik@gmail.com

Abstract

This paper briefly summarizes some recent advances in monocular camera based visual navigation of mobile robots. The first part of this paper describes a self-supervised learning algorithm, which estimates vanishing point of the road and adaptively train color classifiers to detect the road and non-road areas. The second part deals with spatio-temporal consistency for 2D semantic scene analysis, addressed by learning visual similarities between pixels across frames and a simple filtering algorithm in online/causal manner.

Categories and Subject Descriptors

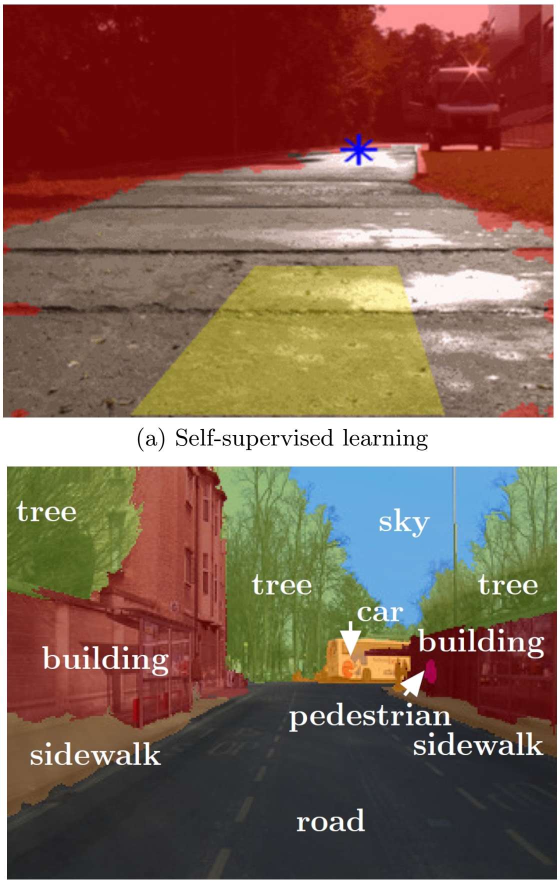
I.4 [**Image Processing and Computer Vision**]: Scene Analysis

Keywords

visual navigation, road detection, semantic scene analysis, spatio-temporal consistency, mobile robots

1. Introduction

During the past few decades, the robotics community has made great efforts in developing autonomous or semi-autonomous robots. Such robots are able to perform desired tasks without continuous human guidance. One of the most fascinating problems for researchers working in the domain of mobile robotics is the development of a robot, which can autonomously operate in structured or unstructured environment. An ultimate goal perfectly represents a project of self-driving cars.


(a) Self-supervised learning

(b) Spatio-temporal consistency

Figure 1: Results of discussed systems.

Reliable perception is crucial for autonomous robots – the goal is to detect drivable surface ahead of the robot and plan the trajectory. This task is not easy even with the most advanced sensors. Usually, common sensors such as laser range finders provide information about obstacles in a near field, however long-range sensing is needed to be able to plan smooth trajectories for high speed vehicles. A combination of short-range sensors with a camera is commonly used to overcome such limitations.

This paper briefly discusses two recently proposed algorithms – while the first one can be used with semi-autonomous robots operating primarily in unstructured environment, the latter can be used with more advanced systems that aim at fully autonomous behaviour.

*Master study programme in field Cybernetics, Automation and Measurement. Supervisor: Dr. Ludek Zalud. Faculty of Electrical Engineering and Communication, Brno University of Technology

Miksik, O. Dynamic Scene Understanding for Mobile Robot Navigation. Information Sciences and Technologies Bulletin of the ACM Slovakia, Special Section on the ACM Student Project of the Year 2012 Competition, Vol. 4, No. 4 (2012) 51-52

2. Self-supervised learning

We demonstrate the demands on algorithms for semi-autonomous robots on the Orpheus-AC reconnaissance mobile robot [5]. Its primary task is to make the measurement and identification in areas with the highest risk of massive contamination. The robot is primarily teleoperated, however, it may be difficult, or even impossible, for the operator to directly control the robot in some situations (signal loss, etc.). For this reason, it might be useful to have a system that would be able to automatically control the robot's movement in order to follow the road, which should be able to: 1) operate under various illumination conditions (direct sunlight, overcast, ...); 2) reliably drive on both high-quality roads as well as on roads barely visible even for humans (sand, concrete, etc.); 3) use a minimum number of sensors – since the robot is intended to work in contaminated areas, it has to be extremely easy-to-decontaminate. The robot is teleoperated, so it is already equipped with a high quality camera, which is an obvious source of data.

By comparison with recently presented state-of-the-art methods, we neither use a laser range finder [1], nor stereo vision [2] for extraction of the training area. Our approach is a fusion of the frequency based estimation of so called vanishing point and probabilistically based texture segmentation. A combination of two different approaches, allows us to solve difficult situations without any a priori knowledge of robot's environment. The key idea of our approach is estimation of the vanishing point, which determines the training area for texture segmentation. Next, road color models are constructed from sample pixels defined by the training area. These models are associated with previously learned models, which are stored in a memory. Further, learned models are adaptively updated. Therefore, the models include both the road colors' history and the current road appearance. A few simple rules define properties of the color segmentation system, like adaptivity speed, selectivity, robustness or behavior in shady and/or overexposed highlighted road segments.

The strategy of our vision system is the following: start with the vanishing point estimation, which is used to detect the training area for self-supervised learning of color models. Next, self-supervised learning continues, however, it is possible to perform road segmentation based on these models. Besides, a combination of two different approaches is advantageous, because in situations like sudden road texture or illumination change, we are still able to estimate the correct course, because if the color models are not consistent with current road surface, it is possible to use a vanishing point until new color models are learned.

3. Spatio-temporal consistency for semantic scene analysis

A semantic scene understanding from images (Fig. 1 b), provides more information about the environment than just road and non-road regions. However state-of-the-art algorithms typically address the problem of scene analysis from a single image [3, 4]. Extending these techniques to temporal sequences of images, as would be seen from a mobile platform, is very challenging. Simply applying the scene analysis algorithm to each image independently is not sufficient because it does not properly enforce consistent labels over time. In practice, the temporally inconsistent predictions result in "flickering" classifications. This effect is not just due to the motion of the camera through the 3D scene: we often observe this behavior even on images of a *static* scene due to subtle illumination changes. These inconsistencies in predictions can have a major impact on robotic tasks in practice, e.g., predicted obstacles may suddenly appear in front of the robot in one frame and then vanish in the next. The situation is further complicated by the need for online, causal algorithms in robotics applications, in which the system does not have access to future frames, unlike video interpretation systems which can proceed in batch mode by using all the available frames.

Our approach is based on the most natural technique for maintaining temporally consistent predictions: a temporal filter based on exponential smoothing over past predictions. Our approach is a meta-algorithm in the sense that it is agnostic to the specific way in which predictions are generated, so that is can be used with any per-frame scene analysis technique. Our only requirement is that the per-frame scene analysis technique predicts a per-pixel label probability distribution instead of a single label.

There exist at least two reasons, why naive averaging cannot work – since we are interested in dynamic scene understanding, the same spatial coordinates in adjacent frames may represent completely different semantic classes. Moreover, only a small portion of pixels in a local neighborhood may correspond to the same semantic class, which is present in the reference pixel. Hence our algorithm consists of two steps: 1) dense optical flow which is able to deal with large displacements is used to initialize a local neighborhood and 2) the past and current predictions are combined by weighted averaging, where the weights correspond to a data-driven learnt visual similarity function, which assigns a high weight between pixels that correspond to each other (and low weight for those that do not) in order to select correct correspondences and accurately propagate predictions over time.

4. Conclusions

This paper briefly discusses two possible approaches to semantic scene understanding of dynamic environments, which is important for navigation of mobile robots, semantic mapping and other applications. The former method is based on a novel combination of vanishing point estimation and color segmentation, the latter improves the stability of semantic predictions in adjacent frames by temporal filtering based on optical flow and weighted averaging. Both approaches proved to be efficient in terms of speed and precision and outperform state-of-the-art algorithms. More details can be found in respective papers.

References

[1] H. Dahlkamp, A. Kaehler, D. Stavens, S. Thrun, and G. R. Bradski. Self-supervised monocular road detection in desert terrain. In *Robotics: Science and Systems*, 2006.

[2] T.-C. Dong-Si, D. Guo, C. H. Yan, and S. H. Ong. Robust extraction of shady roads for vision-based ugv navigation. In *IROS*, 2008.

[3] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr. What, where & how many? combining object detectors and crfs. In *ECCV*, 2010.

[4] D. Munoz, J. A. Bagnell, and M. Hebert. Stacked hierarchical labeling. In *ECCV*, 2010.

[5] L. Zalud. Orpheus - reconnaissance teleoperated robotic system. In *16th IFAC World Congress, Prague, Czech Republic*, 2005.