International Journal of Computer Vision (2023) 131:407–430
https://doi.org/10.1007/s11263-022-01706-5

Robots Understanding Contextual Information in Human-Centered Environments Using Weakly Supervised Mask Data Distillation

Daniel Dworakowski[1] · **Angus Fung**[1] · **Goldie Nejat**[1]

Received: 23 November 2020 / Accepted: 18 October 2022 / Published online: 11 November 2022

Abstract
Contextual information contained within human environments, such as text on signs, symbols and objects provide important information for robots to use for exploration and navigation. To identify and segment contextual information from images obtained in these environments data-driven methods such as Convolutional Neural Networks (CNNs) can be used. However, these methods require significant amounts of human labeled data which is time-consuming to obtain. In this paper, we present the novel Weakly Supervised Mask Data Distillation (WeSuperMaDD) architecture for autonomously generating pseudo segmentation labels (PSLs) using CNNs not specifically trained for the task of text segmentation, e.g., CNNs alternatively trained for: object classification or image captioning. WeSuperMaDD is uniquely able to generate PSLs using learned image features from datasets that are sparse and with limited diversity, which are common in robot navigation tasks in human-centred environments (i.e., malls, stores). Our proposed architecture uses a new mask refinement system which automatically searches for the PSL with the fewest foreground pixels that satisfies cost constraints. This removes the need for handcrafted heuristic rules. Extensive experiments were conducted to validate the performance of WeSuperMaDD in generating PSLs for datasets containing text of various scales, fonts, orientations, curvatures, and perspectives in several indoor/outdoor environments. A detailed comparison study conducted with existing approaches found a significant improvement in PSL quality. Furthermore, an instance segmentation CNN trained using the WeSuperMaDD architecture achieved measurable improvements in accuracy when compared to an instance segmentation CNN trained with Naïve PSLs. We also found our method to have comparable performance to existing text detection methods.

Keywords Weakly supervised learning for robots · Environment context identification · Segmentation and labeling · Robot navigation and exploration

1 Introduction

Human-centered environments contain an abundance of contextual information such as text on signs, symbols, and objects that are used as landmarks to help guide users with point-to-point navigation in unknown environments (Vilar et al., 2014), and update maps of the environment (Peng et al., 2018). Service robots working in varying human-centered (Dworakowski et al., 2021) environments can exploit these types of contextual information to aid with navigation. For example, robots can use text on aisle signs in grocery stores to determine which aisles to search for a particular item (Thompson et al., 2018). They have also used contextual information for mapping and localization. Namely by using an annotated map of an office with room placards for goal directed navigation (Case et al., 2011). Robots have also created semantic maps using product locations (Cleveland et al., 2017), maps from unique text landmarks identified in images (Wang et al., 2015), and have used salient objects identified from learned features (e.g., edges, contours, etc.) for visual

Communicated by Frederic Jurie.

✉ Daniel Dworakowski
daniel.dworakowski@mail.utoronto.ca

Angus Fung
angus.fung@mail.utoronto.ca

Goldie Nejat
nejat@mie.utoronto.ca

[1] Autonomous Systems and Biomechatronics Laboratory (ASBLab), Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, ON M5S 3G8, Canada