Utilization of 2.5D imagery for AI-based object detection

PROJECT PROPOSAL

By BENEDICT THEKKEL
45813452
Commenced: Semester 2, 2023
Mode of study: Full-time - Internal
Supervisors: Brian Lovell

School of Electrical Engineering and Computer Science
The University of Queensland.

Contents

Abstract

The objective of this project is to evaluate the transformation from a conventional 2D image to a 2.5D image obtained from a singular source. This integration entails harnessing a depth estimation model alongside AI object detection techniques. If accurate depth information can be calculated through the implementation of depth estimator models using a single camera source, and if the utilization of 2.5D image input yields improved and more precise image segmentation, it follows logically that the combination of these two techniques would yield a synergistic enhancement in image segmentation compared to the utilization of original 2D images. In the context of this project, the central problem to address revolves around determining the advantages and improvements that can be achieved through the fusion of depth estimator models and 2.5D image segmentation models, when contrasted with the conventional practice of image segmentation solely using 2D images.

The relevance of research in this field extends across industries reliant on object detection and image segmentation, encompassing domains such as augmented reality (AR), virtual reality (VR), 3D modeling, medical imaging, autonomous driving, and facial recognition. These technologies hold promise for transformative advancements, enabling tasks ranging from immersive digital experiences to precise medical diagnoses and safer autonomous vehicles. Such applications underscore the relevance of this research across a multitude of sectors, promising to revolutionize various industries and domains.

1 Introduction

1.1 Topic

Significant progress in Convolutional Neural Networks (CNN) has considerably enhanced the efficacy of semantic segmentation [1]. The proliferation of range sensors like Kinect has led to the widespread availability of depth data. Consequently, there is a growing enthusiasm for harnessing depth information to enhance semantic segmentation. Depth data introduces geometric details that are absent in color channels, thereby enhancing the representation acquired through CNN. The advantages of employing 3D imagery for object segmentation and detection have been extensively substantiated. [2].

1.2 Scope of Previous Research

Prior investigations have addressed the utilization of 2.5D imagery for object segmentation employing diverse techniques, including R-CNN [3]. Research in the realm of depth estimation has progressed from combining RGB with depth, stereo images, and lidar to monocular sequences, often utilizing datasets like KITTI [4] . The generation of 2.5D imagery from a solitary source has already demonstrated remarkable achievements [5] [6]. Additionally, deep learning approaches have achieved precise depth estimation through techniques such as the Laplacian pyramid [7].

1.3 Goals

This project seeks to combine depth estimation and 2.5D image generation technologies. It aims to evaluate the shift from conventional 2D images to richer 2.5D representations sourced from a single camera. The goal is to reveal the benefits of this integration, enhancing image analysis, object detection, and segmentation for various applications.

1.4 Relevance

The research in this field holds great significance for industries reliant on object detection and image segmentation. Its applications span across sectors such as AR, VR, 3D modeling, medical imaging, autonomous driving, and facial recognition, promising transformative advancements. As exemplified by Nebiker et al. [2], this technology's diverse potential underscores its relevance across industries, offering improved capabilities in tasks ranging from immersive digital experiences to accurate medical diagnoses and safer autonomous vehicles.

2 Background and Literature Review

2.1 Literature Review

2.1.1 Depth Estimators

Early research primarily employed statistical features from color images to estimate monocular depth. Torralba and Oliva [8] explored spectral magnitude properties linked to depth changes. Saxena et al. [9] used a planar layout approach, incorporating 3D position and orientation via Markov random field (MRF) analysis with features like edge orientations and chromatic values. Chun et al. [10] estimated depth maps for single indoor scenes using ground region position data, especially relative distance from the highest floor point. Recent trends focus on identifying suitable depth values by comparing structural similarity with scenes possessing real depth information. Karsch et al. [11] proposed a method assessing spectral coefficient similarity, refining results with techniques like SIFT flow [12] warping. [13] tackled optimal depth selection through coarse-to-fine cluster-based learning. Yet, patch-based aggregation may struggle to represent intricate geometries, resulting in blurred estimations.

2.1.2 2.5D image segmentation

Attention-driven segmentation is characterized by a two-stage process: the initial stage focuses on identifying candidate object locations, while the subsequent stage involves the segmentation of objects. Various attention-driven segmentation approaches, exemplified by [14], [15], and [16], offer diverse solutions for both stages, often tailored to specific search tasks in robotic contexts. Mishra et al. [15] introduced an active segmentation framework that utilizes fixation points to define object boundaries, advocating for the utilization of visual attention mechanisms in the selection process. In a related vein, [16] extended the active segmentation concept by incorporating "simple objects and border ownership," employing depth, color, and motion information, and introducing a novel fixation point calculation approach. Kootstra et al. [14] contributed to the field by proposing an attention-driven graph-cut segmentation technique. This method localizes objects using fixation points derived from 2D symmetry saliency maps and applies energy minimization based on depth, color, and support plane constraints. However, the segmentation algorithms found in [14] and [16] are specialized for table-top scenarios and rely on support plane data, limiting their suitability for intricate, cluttered, and occluded scenes. Fixation point selection commonly involves using fixations as attention cues within saliency maps, as observed in both [14] and [17]. In the pursuit of enhanced fixation point extraction, Potapova et al. [17] proposed 2.5D symmetry-based saliency maps, revealing their superior capacity to capture scene properties compared to conventional 2D saliency maps.

2.2 Research Problem

If precise depth can be computed through depth estimator models utilizing a single camera source, and the utilization of 2.5D image input enhances and refines image segmentation accuracy, it follows that the amalgamation of these techniques could lead to superior image segmentation outcomes compared to conventional 2D images. In the context of this project, the task revolves around discerning the benefits and enhancements achievable through the fusion of depth estimator models and 2.5D image segmentation models, in contrast to direct image segmentation using 2D images.

3 Program and Design of the Research Investigation

3.1 Objectives

This project aims to enhance object detection by merging a depth estimation model with AI techniques. This integration leverages both technologies to improve accuracy, especially in complex scenarios. The goal is to create a more robust and adaptable object detection system, advancing computer vision capabilities.

3.2 Methodology

- **Dataset Construction:** The project begins with the acquisition of object images to create a dataset similar to KITTI, serving as a benchmark for the research. This dataset will be pivotal for training and evaluating the depth estimation and object detection models.
- **Depth Estimation Model Development:** A crucial aspect of the project involves the development of a sophisticated depth estimation model. This model will be specifically designed to accurately predict depth information from single-camera images, contributing to the creation of 2.5D imagery.
- **Integration of AI Object Detection:** The project seamlessly integrates cutting-edge AI object detection techniques, utilizing a FASTAI object detection model. This model will be applied to both the original 2D images and the enhanced 2.5D images generated through depth estimation.
- **Comparative Analysis:** The final step involves a comprehensive comparative analysis of the outcomes. The performance of the object detection model applied to the original images will be compared with its performance on the 2.5D images. This analysis will provide valuable insights into the efficacy and enhancements achieved through the fusion of depth information.

3.3 Research Plan

This research initiative involves a comprehensive exploration of prior literature pertaining to the utilization of 2.5D or 3D imagery for image segmentation, delving into the methodologies and findings of earlier studies. Concurrently, a meticulous examination of research concerning models designed to generate depth estimations from monocular sources will be undertaken. The objective is to synthesize this accumulated knowledge and expertise in order to formulate an AI model capable of seamlessly integrating both functionalities – leveraging a solitary camera source for achieving the combined capabilities of 2.5D/3D image segmentation and accurate depth estimation.

3.4 Resources and Funding Required

This project would require a camera and hardware application for demonstrating the proof of concept.

3.5 Milestones

Assessment Task	Milestones	Weighting
Online Quiz UQ Academic Integrity Tutorial	17/08/2023 15:00	Passed
Report Proposal Draft	24/08/2023 15:00	Pass/Fail
Research Proposal Project Proposal	21/09/2023 15:00	10%
Seminar	09 Oct 23 - 13 Oct 23	15%
Progress Seminar Participation Seminar Participation	13-Oct-23	Pass/Fail
Demonstration Poster and Demonstration	17-May-24 no later than	25%
Thesis Thesis Report Submission	3/06/2024 15:00	50%

3.6 OHS Risk and Ethics assessment

3.6.1 Risk

The tasks encompassed by this project entail image data collection, the development of machine learning algorithms, and the creation of deep learning models. These activities can all be carried out within low-risk laboratory environments, adhering to the standard laboratory safety protocols governed by occupational health and safety guidelines.

3.6.2 Ethics

The datasets utilized in this project comprise publicly accessible and freely available academic resources.

4 References

References

[1] Y. Xing, J. Wang, X. Chen, and G. Zeng. 2.5 d convolution for rgb-d semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1410–1414. IEEE, 2019.

[2] S. Nebiker, J. Meyer, S. Blaser, M. Ammann, and S. Rhyner. Outdoor mobile mapping and ai-based 3d object detection with low-cost rgb-d cameras: The use case of on-street parking statistics. *Remote Sensing*, 13(16):3099, 2021.

[3] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 345–360. Springer, 2014.

[4] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] M. Song, S. Lim, and W. Kim. Monocular depth estimation using laplacian pyramid-based depth residuals. *IEEE transactions on circuits and systems for video technology*, 31(11):4381–4393, 2021.

[6] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

[7] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian. Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*, 63(9):1612–1627, 2020.

[8] A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9):1226–1238, 2002.

[9] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5):824–840, 2009.

[10] C. Chun, D. Park, W. Kim, and C. Kim. Floor detection based depth estimation from a single indoor scene. In *2013 IEEE International Conference on Image Processing*, pages 3358–3362, 2013.

[11] K. Karsch, C. Liu, and S. B. Kang. Depth extraction from video using non-parametric sampling. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 775–788. Springer, 2012.

[12] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011.

[13] J. L. Herrera, C. R. del Blanco, and N. García. Automatic depth extraction from 2d images using a cluster-based learning framework. *IEEE Transactions on Image Processing*, 27(7):3288–3299, 2018.

[14] G. Kootstra, N. Bergström, and D. Kragic. Fast and automatic detection and segmentation of unknown objects. In *2010 10th IEEE-RAS International Conference on Humanoid Robots*, pages 442–447, 2010.

[15] A. Mishra, Y. Aloimonos, and C. L. Fah. Active segmentation with fixation. In *2009 IEEE 12th International Conference on Computer Vision*, pages 468–475, 2009.

[16] A. K. Mishra. Visual segmentation of simple objects for robots. *Robotics: Science and Systems (RSS)*, 2011.

[17] E. Potapova, M. Zillich, and M. Vincze. Local 3d symmetry for visual saliency in 2.5 d point clouds. In *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11*, pages 434–445. Springer, 2013.

5 Appendix