Exploration of Space Discretization Models for Single Image Depth Estimation

ANDREA NARDELLI TIANZE WANG

andnar | tianzew @kth.se

30th November 2018

Abstract

This work aims to explore the use of different discretization functions in an ordinal regression framework for monocular depth estimation of a single image. Existing work in this field uses uniform space discretizations without taking into account relative error and without weighting under- or overestimation differently, an addition that is in our opinion application-critical in a practical setting. Our goal is hence exploring these discretizations for the purpose of improving the accuracy of depth estimation, and investing how the ordinal regression framework can be applied to other depth estimation models.

Contents

1 Introduction

1.1 Aim & Objective

Depth estimation is the process of predicting the depth map of a 2D image or video. While current methods for estimating depth from videos or images from dual cameras, depth estimation performance in the monocular vision are still to be improved. The aim of the research is to improve the performance of monocular depth estimation from a single image (abbr. as MDE hereafter).

The objective of the research is to find non-uniform discretization spaces that can be used to increase the accuracy of prediction for MDE problems, which in turn can boost the performance of 3D space mapping in fields such as robotics navigation, self-driving cars, human pose estimation and more.

1.2 Goals

The first goal is identifying non-uniform discretization strategies that can improve the performance of [1] in the examined datasets. While the logarithm-based strategies used in [1] shows performance improvement, it would also be interesting to see if other strategies would work and what might be the key factors and rationale for choosing a particular transformation.

A secondary goal of the project is investigating how the ordinal regression and non-uniform discretization strategies affect the performance of other MDE models. As shown in [1], ordinal regression which "aims to learn a rule to predict labels from an ordinal scale" has demonstrated an advantage in MDE models.

2 Background

Depth information can be estimated from stereo images or motion sequences, which have relatively rich information for understanding the 3D structures of the scene. However, estimating depths from a single image is yet another challenge for the simple reason that depth remains uncertain when only local image features are given. Various monocular cues, like texture variations and gradients, defocus, color/haze, etc, which contain some depth related information can also offer crucial support to depth estimations [2]. Yet another problem is that these cues might not be available in some datasets.

2.1 Theoretical Framework

An image is formalized as a real-valued matrix I with dimensions $M \times N$. The depth map D_I represents the depth map of image I with the same dimensions and represents the "ground truth" of depth map estimation. The goal of depth estimation is creating a function f which outputs $\hat{D}_I$, an approximation of D_I with dimensions $P \times Q$ where $P \leq M$ and $Q \leq N$. We can hence define f as

$$f(I) : \mathbb{R}^{M \times N} \to \mathbb{R}^{P \times Q}.$$

Note that the estimated depth map can be smaller than the ground truth: this is sometimes done in order to increase the speed of computation and model training. Fu et al. [1] introduce an ordinal regression framework in which the task of approximating the depth is transformed to predicting an ordinal class out of K classes for each element in $\hat{D}_I$. In other words, each classes corresponds to a depth range and each element in $\hat{D}_I$ belongs to one of these classes. As mentioned, this transformation to an ordinal regression is achieved through a logarithmic transformation which is formalized as

$$t_i = e^{\log(\alpha) + \frac{i \cdot \log(\beta/\alpha)}{K}}$$

where $[\alpha, \beta]$ represent the original regression range and $t_i \in \{t_0, t_1, \ldots, t_K\}$ are the discretization thresholds.

2.2 Literature Review

In this subsection, some of the state-of-the-art methods for depth estimation of a single image will be reviewed.

Eigen et al. [3] present a new method that address the problem by using two deep network stacks where one of them makes a coarse global prediction based on the entire image, and the other refines this prediction locally. They also use scale-invariant errors to measure depth relations rather than using scale.

He et al. [4] demonstrate experimentally that focal length has a great influence on accurate depth recovery. They also propose a new deep neural network to estimate depth through effectively fusing the middle-level information on the fixed focal-length dataset which outperforms the state-of-the-art methods developed upon pretrained VGG.

Zhang et al. [5] aim to solve the problem that commodity-level depth cameras often fail to capture depth information for shiny, bright, transparent, and distant surfaces. They have proposed a deep network that takes an RGB image and predicts dense surface normals and occlusion boundaries. Then these estimations are combined together with the raw depth information provided by the camera to generate the depths for all pixels of the image including the missing ones from the camera. They have also shown experimentally that their proposed approach has better depth completions than its counterparts.

Qi et al. [6] propose Geometric Neural Network (GeoNet), which is built on top of two-stream CNNs, to jointly predict depth and surface normal maps from a single image. The key contribution of the model is incorporating geometric relation between depth and surface normal via their new depth-to-normal and normal-to-depth networks. Experiments have shown that these two networks make the underlying model to efficiently estimate depth and surface normal in a high consistency manner and achieves top performance for surface normal estimation and state-of-the-art accuracy for depth estimation.

Lee et al. [7] propose a deep learning algorithm for single-image depth estimation based on Fourier frequency domain analysis. Apart from a convolutional neural network, they also propose a depth-balanced Euclidean loss which is reliable for training a wide variety of networks for depths estimations. They also take advantage of complementary properties of small and large ratio cropped images by combining the multiple candidates in the frequency domain. Experiments have shown that the proposed algorithm can achieve state-of-the-art performance.

Atapour-Abarghouei et al. [8] observe that while monocular depth estimation via learning-based approaches yields promising results, most of them rely on large volumes of ground truth depth data or predict disparity as an intermediary step with a secondary supervisory signal. To solve this, they introduce image style transfer and adversarial training to solve the domain bias brought by pixel-perfect synthetic data. Experiment results have shown that their approach is comparable to state-of-the-art techniques.

3 Research Questions & Hypotheses

3.1 Problem statement

Previous evaluation measures for MDE did not consider the relative error of an estimation compared to its ground truth, which can result in depth maps that are inaccurate for objects close to the camera. For example, the relative error between a depth prediction of 2m with ground truth 1m is 100%, whereas a prediction of 101m with ground truth 100m is 1%.

The main contribution of [1] consists of using a discretization to transform the output space into an ordinal regression in a logarithmic space which takes into account relative error. This project will investigate different strategies/transformations in order to implement this discretization. In addition, it will explore the inverse transformation (i.e. from the discretized output space back to the original regression) for improved performance evaluation.

3.2 Problem

Except for the contribution in [1], the current problem with MDE evaluation measures is that they do not take into account the relative error of a prediction with regards to its ground truth.

3.3 Hypothesis

The use of non-uniform discretization spaces can improve the performance of existing solutions of MDE problems and can specialize a model for particular applications. The ordinal regression framework can be applied as a general approach in MDE models.

4 Research Methodology

4.1 Organization

The project will be conducted in a two-person group, expanding upon the work previously accepted in the 2018 edition of CVPR (Conference on Computer Vision and Pattern Recognition).

4.2 Allocation of responsibilities

Both team members are responsible for writing the project plan and report, in addition to researching existing bibliography connected to the main paper we want to expand upon.

In particular, Tianze Wang is responsible for managing the computing resources, contacting original authors if needed, obtaining and preprocessing the datasets.

Andrea Nardelli is responsible for implementing the models and presenting the work.

4.3 Method

The research will use an experimental methodology by testing different discretization strategies and evaluating their results. The efficacy of non-uniform discretization will be empirically tested on other MDE models.

The performance of the models will be analyzed quantitatively using standard MDE metrics, whereas the output depth maps as images will be analyzed qualitatively.

4.4 Tasks & Procedures

Due to the fact that the storage and computation needed by this research goes beyond our current capabilities, we asked the makers of Hops* for support which they generously provided us. All the experiments will be A virtual machine will be prepared to run the experiments, including preprocessing the image datasets.

The MDE model mentioned in [1] together with a number of different discretization strategies will be re-implemented with the PyTorch framework, as the original Caffe implementation with noninternal packages is not available. The different discretization strategies will be evaluated on the same public datasets used in [1]. Results of the experiments will be analyzed and a comparison across different strategies will be made through a quantitative and qualitative analysis of the output depth maps. Lastly, the ordinal regression and non-uniform discretization approach will be evaluated on other MDE models.

A report detailing all of the work will be provided.

4.5 Data Collection

The data comes from publicly available datasets used in the field of computer vision. In particular as we aim to reproduce the work of [1], the datasets used are KITTI [9], Make3D [10, 11], NYU Depth v2 [12].

* https://www.hops.io/

5 Expected Outcomes

Apart from the discretization strategies mentioned in [1], we are expecting that we will find one or more other discretization strategies that will also improve the performance of the depth estimation model.

To be more specific, we have observed that in certain applications, e.g. autonomous driving, a false negative where the model estimates a near point to be far away is more of a serious problem than a false positive where a far away point is estimated to be near the vehicle. While the former case might lead to serious accident, the latter case will cost no more than the vehicle to stop. Thus, we would also like to explore if there is any discretization strategy that punish false negatives more than false positives.

Apart from that, we are also interested in finding out the role of ordinal regression in depth estimation model.

6 Milestones

The project started on the 19th September 2018 and will end on the 15th January 2019. The following list covers the milestones of our project both in terms of deliverables (proposals, plans, reports) and in terms of work plan. Already achieved milestones are in bold.

1. **19th September 2018**: First draft of project proposal is submitted.

2. **21st September 2018**: Presentation and peer review of ethics & sustainability concerns for the proposed project are submitted.

3. 12th October 2018: First draft of research plan and presentation is submitted for peer review.

4. 15th October 2018: The data from the public datasets is obtained and preprocessing pipeline is written and executed, resulting in the preprocessed data.

5. 29th October 2018: The original model is re-implemented as described in the previous sections.

6. 5th November 2018: Particular care is given to reproduce the results presented in [1], in order to identify possible errors in our implementation. Simultaneously, different discretization strategies will be implemented as described in Section 4.

7. 19th November 2018: Testing of the examined discretization strategies in different MDE models.

8. 30th November 2018: First draft of research report and presentation is submitted for peer review.

9. 3rd December 2018: In parallel to completing the last batch of analysis, start work on collecting results and incorporate them into the draft of the report.

10. 10th December 2018: Peer review & opposition on latest report draft.

11. 17th December 2018: Code freeze.

12. 11th January 2019: By this date, final seminar and opposition with peer review will be completed.

13. 15th January 2019: Turn-in of final project report.

7 Risks & Ethics and Sustainability

In the field of computer vision, ethical concerns are a major source of discussion. Our research in depth estimation take into account these aspects. The most common controversial application regards use of computer vision on images from CCTV cameras in a kind of tradeoff of security vs privacy. Whereas on one side applying computer vision may help in providing additional security by e.g. tracking criminal perpetrators, on the other side it may result in an infringement of privacy for people not connected to the crime. Similar concerns arise when considering cameras in public places and how these can be used for surveillance. In particular our research into depth estimation could improve existing technologies for tracking individuals with these security cameras or even in espionage situations in which an inaccessible space is 3D mapped to reconstruct a normally inaccessible building.

In an example application of depth estimation for autonomous driving, our research must considers the risks of inaccurate predictions. For example, a depth prediction that is larger than what the actual depth is (we can think of it as a "false positive") may result in an autonomous car thinking it has more space to navigate than what it actually has. For this application, this can be argued to be worse than a "false negative" in which a smaller than reality prediction is created by the system.

Given the risks mentioned above, one might ask if it is ethical to release our work as open-source, which makes it easily available not only to the ethical user but also to users with negative intents in mind. Indeed this is a major concern for all research and its possible applications, and in the case of open-source one could argue that by publishing our code/implementation, nefarious applications can be created with very little effort. We have two counter points to this:

1. Ethical applications that make use of our work are similarly easier to create and contribute back to research in a virtuous loop, creating a positive environment where others can build on our work inspired by the Sustainable Development Goal #17[*] which mentions "strengthening the means of implementation and revitalise the global partnership for sustainable development".

2. As history reminds us, nefarious applications of research are always a possibility. We personally believe that an ethical reflection is mandatory by users of our work, we condemn any non-ethical use of our work or use which harms individuals or their privacy.

This work also aims to contribute to Sustainable Development Goal #8[*] which reads "promoting sustained, inclusive and sustainable economic growth, full and productive employment, and decent work for all" by contributing to research that allows the automation of tedious, manual and excruciating jobs and the requalification of employees for better working & life conditions.

[*] https://sustainabledevelopment.un.org/

References

[1] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao, "Deep Ordinal Regression Network for Monocular Depth Estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, United States, 2018, pp. 2002–2011. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/papers/Fu_Deep_Ordinal_Regression_CVPR_2018_paper.pdf

[2] Saloni Bahadur, Rashmee Shrestha, Yalapi Sumaharshini, Gnv Ravi Teja, and Kalpitha.N, "Literature Review on Various Depth Estimation Methods for an Image," *International Journal of Research - Granthaalayah*, vol. 5, no. 4, pp. 8–13, Apr. 2017. doi: https://doi.org/10.5281/zenodo.572287. [Online]. Available: http://granthaalayah.com/Articles/RACSIT/IJRG17_RACSIT_02.pdf

[3] David Eigen, Christian Puhrsch, and Rob Fergus, "Depth Map Prediction from a Single Image using a Multi-Scale Deep Network," in *Advances in neural information processing systems*, vol. 27. Montréal, Canada: Curran Associates, Inc., 2014, pp. 2366–2374. [Online]. Available: http://papers.nips.cc/paper/5539-depth-map-prediction-from-a-single-image-using-a-multi-scale-deep-network.pdf

[4] Lei He, Guanghui Wang, and Zhanyi Hu, "Learning Depth from Single Images with Deep Neural Network Embedding Focal Length," *CoRR*, vol. abs/1803.10039, 2018. doi: 10.1109/TIP.2018.2832296. [Online]. Available: http://arxiv.org/abs/1803.10039

[5] Yinda Zhang and Thomas Funkhouser, "Deep Depth Completion of a Single RGB-D Image," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, United States, Jun. 2018, pp. 175–185. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/papers/Zhang_Deep_Depth_Completion_CVPR_2018_paper.pdf

[6] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia, "GeoNet: Geometric Neural Network for Joint Depth and Surface Normal Estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, United States, Jun. 2018, pp. 283–291. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/papers/Qi_GeoNet_Geometric_Neural_CVPR_2018_paper.pdf

[7] Jae-Han Lee, Minhyeok Heo, Kyung-Rae Kim, and Chang-Su Kim, "Single-Image Depth Estimation Based on Fourier Domain Analysis," in *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, United States, Jun. 2018, pp. 330–339. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/papers/Lee_Single-Image_Depth_Estimation_CVPR_2018_paper.pdf

[8] Amir Atapour-Abarghouei and Toby P. Breckon, "Real-Time Monocular Depth Estimation using Synthetic Data with Domain Adaptation via Image Style Transfer," in *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, United States, Jun. 2018. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/papers/Atapour-Abarghouei_Real-Time_Monocular_Depth_CVPR_2018_paper.pdf

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013. doi: 10.1177/0278364913491297. [Online]. Available: https://doi.org/10.1177/0278364913491297

[10] Ashutosh Saxena, Sung H. Chung, and Andrew Y. Ng, "Learning depth from single monocular images," in *Advances in neural information processing systems*, Hyatt Regency Vancouver, Vancouver, British Columbia, Canada, 2006, pp. 1161–1168. [Online]. Available: http://papers.nips.cc/paper/2921-learning-depth-from-single-monocular-images.pdf

[11] Ashutosh Saxena, Min Sun, and Andrew Y. Ng, “Make3d: Learning 3d scene structure from a single still image,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 824–840, May 2009. doi: 10.1109/TPAMI.2008.132. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4531745

[12] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in *Computer Vision – ECCV 2012*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. ISBN 978-3-642-33715-4 pp. 746–760.