Embedded Selforganizing Systems (Vol 7. No 2. 2020) (pp. 4-7)

Embedded Selforganizing Systems

Issue Topic: "Towards Robust Situation Awareness in Autonomous Vehicles"

Real-time 3D Perception of Scenes with Monocular Camera

Shadi Saleh
Department of Software Engineering
Technische Universität Chemnitz
Chemnitz, Germany
shadi.saleh@informatik.tu-chemnitz.de

Shanmugapriyan Manoharan
Embedded Systems
Technische Universität Chemnitz
Chemnitz, Germany
shanmugapriyan.manoharan@s2017.tu-chemnitz.de

Wolfram Hardt
Department of Computer Engineering
Technische Universität Chemnitz
Chemnitz, Germany
wolfram.hardt@informatik.tu-chemnitz.de

Abstract— Depth is a vital prerequisite for the fulfillment of various tasks such as perception, navigation, and planning. Estimating depth using only a single image is a challenging task since the analytic mapping is not available between the intensity image and its depth where the features cue of the context is usually absent in the single image. Furthermore, most current researchers rely on the supervised Learning approach to handle depth estimation. Therefore, the demand for recorded ground truth depth is important at the training time, which is actually tricky and costly. This study presents two approaches (unsupervised learning and semi-supervised learning) to learn the depth information using only a single RGB-image. The main objective of depth estimation is to extract a representation of the spatial structure of the environment and to restore the 3D shape and visual appearance of objects in imagery.

Keywords— Depth estimation single image, unsupervised learning, semi-supervised learning, realt time.

I. INTRODUCTION

Understanding of visual scenes is a vital component of many applications of Artificial Intelligence, ranging from autonomous vehicles to the navigation of household robots and even automatic annotation of imagery for the blind. The real-time and reliable extraction of high-level semantic information from the visual world is the crucial factor for the safety and correctness of these critical tasks such as pose estimation, recognition task, semantic labeling, and 3D Object Detection [1, 2, 3]. There are many potential advantages that could be obtained by gaining more knowledge about the 3D geometry of scenes, for example, integrated object detection with distance estimation [4]. Moreover, the majority of obstacle detection and avoidance methodologies are based on predefined object lists, in which the model is trained to detect specific object types within the scene, such as pedestrians and vehicles. The above does not, however, hold for different object types such as debris and roadside obstacles that may appear in the scene. That means the model is not able to detect objects in case of unidentified obstacles, misclassification, or object occlusion situations. Therefore, accurate depth perception for moving and stationary obstacles can provide more knowledge about the environment, which helps autonomous vehicles to take valid action in critical situations based on 3D perception information.

This problem in stereo vision sensors is resolved by computing a disparity image from the stereo image pair to extract depth information as described [8]. The alternative solution for obtaining the depth would be to employ range sensors such as Lidar or radar. These are naturally highly accurate sensors that provide highly precise depth measurements. However, these sensors did not provide high-resolution information and are more expensive than a traditional monocular camera and therefore not very common in the average consumer vehicle. In fact, working on the depth estimation and particularly in the application of the autonomous vehicle, it is a real challenging task due to several factors such as occlusion, the dynamic object in the scene, and imperfect stereo correspondence. Furthermore, at high levels, reflective, transparent, mirroring surfaces are the major enemy of the stereo matching algorithm. For instance, the windshield of a vehicle usually decreases the matching and therefore the quality of the estimated depth.

Based on depth information the system is able to perform the proper procedures to avoid collisions with obstacles that are determined to be too close. Therefore, by employing more robust depth estimation from a monocular camera, the advantages of active safety systems could be used by a larger segment of the vehicle fleet. In addition, most studies rely on labelled data for depth estimation. In this study, an adaptive deep-learning approach will be used to bridge the gap between labelled and unlabeled data, and thus, only queries the samples that would lead to increase the accuracy. The objective is to

reduce model complexity and consequently the number of samples required for training in order to maximize the accuracy and speed of deep learning algorithms.

II. RESEARCH QUESTIONS AND OBJECTIVES

Depth perception refers to the ability of our eye and brain to add a third dimension. Human is able to easily perceive a 3D structure of scene from a single 2D image. Starting from how people generally perceive depth, this will provide us with valuable insights into estimating depth, as most of these methods (for estimating depth) are derived from our human visual system. Both machine vision and human perception share similarities in the way the image is formed. In theory, light beams that strike surfaces from a source are reflected and redirected back to our retina, where they are projected, and our eye then handles them as 2D, same as how the camera image is formed on a 2D image plane.

As illustrated in Figure1, we could easily make the assumption that the gray car is closer to the camera than the white car and that the white car is closer to the camera than the building in the background. We would also be able to recognize that the road is flat or facing the sky. Moreover, we can even observe that the gray car has a cuboid shape without looking at its hidden surfaces. This is a remarkable talent, as a single 3D image is in itself a quite ambiguous task because 2D image could be the projection from an infinite number of different 3D perspectives, therefore, we have to depend on our visual experience to figure out how to resolve this ambiguity. Our eyes are able to perceive 2D projections of the 3D world as we move and view the 3D world from different viewpoints. However, we have learned, after observing millions of 2D projections, how to infer 3D from a single image. The mechanism at work here is that our brain begins to process the input visual signals by recognizing patterns such as size, texture, and motion around the scene referred to as depth cues. There is no depth information about the image, however, we could somehow easily analyze and reconstruct the depth information. We are able to perceive which aspects of the image are close and which are further away from us. Furthermore, these depth cues allow us to observe objects and surfaces that are assumed to be in flat images as 3D scene.

In this study, we follow a similar approach and try to learn 3D perception of scene from a large number of 2D views with a few labeled ground truths. Our main objective is to predict the 3D perception of a scene represented as a depth image, acquired by a monocular camera using edge computing hardware, in order to estimate a paradigm/model of the 3D

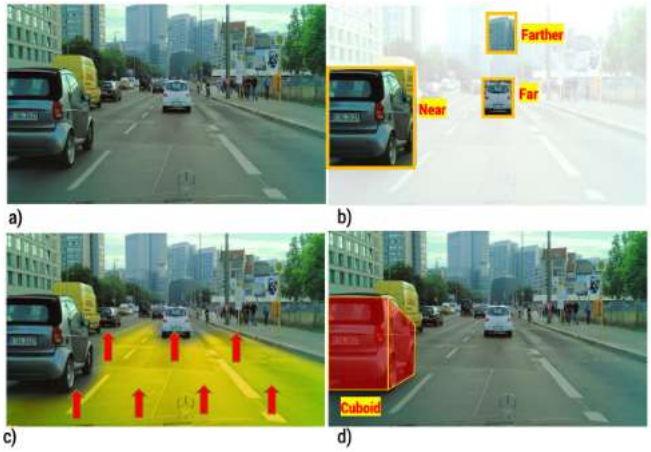

Figure1: perceiving 3D sense from 2D image.

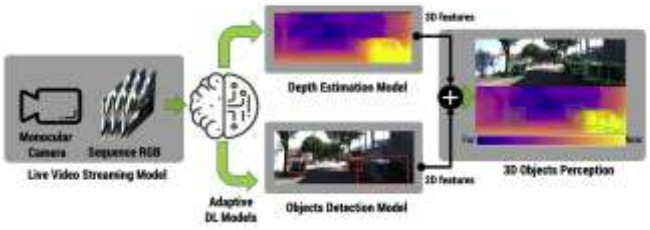

Figure2: Proposed Approach for 3D Perception of Scene with Monocular Camera

geometry of the scene for estimation of freely navigable space with minimal space and power consumption requirements as well as real-time capability. Specifically, we address the following research question: *Is it feasible to retrieve dense, complete depth image information with 3D object detection from a single RGB frame using edge computing hardware with real-time capability?*

III. PROPOSED SOLUTION APPROACH

Our approach will employ a highly efficient live video streaming from a monocular camera and state-of-the-art adaptive depth learning approach for 3D perceptual object recognition by integrated depth estimation with object recognition, which is able to perceive 3D spatial information from the environment in real-time. where adaptive depth learning integrates the previous developments of rule-based, primitive machine learning and deep learning strategies for machine intelligence to provide more reliable, faster and more easily interpretable models, as shown in Figure 2. Our proposed approach would implement two adaptive deep learning models. In the first model, a sequence of RGB images acquired by a monocular camera using a live streaming video model is fed as input into a robust depth estimation model where the relevant depth is estimated, and then the predicted depth information helps us to retrieve the 3D features. On the other hand, the second model integrates detected objects with depth information to estimate and 3D object perception.

IV. DEPTH ESTIMATION TECHNIQUES

Estimating depth using only a single RGB frame is often considered an ill-posed and inherently ambiguous challenge, because the analytic mapping is not available between the intensity image and its depth where the features cue of the context is usually absent in the single image. Traditionally in computer vision, the depth is estimated from 2 well-known methodologies. Namely:

- Depth derived from monocular images (either static or sequential images).
- Depth of stereoscopic images by utilizing the epipolar geometry.

The estimated depth on the basis of visual sensors has gained much in attraction and outstanding outcomes since the introduction of depth learning. A considerable number of ongoing researches has been carried out to solve these challenges. Deep learning approach has proven its outstanding performance in high-level perceptual and cognitive tasks such as detection, detection, and understanding of scenery. Depth

natural approach forward. Currently, there are 3 broad, deep learning frameworks for obtaining depth information:

Supervised Learning Depth Estimation: Estimation of learning depth under supervision: In this approach, the depth is estimated directly from monocular 2D images via a labeled dataset by reducing regression loss function. Since then, various approaches have been developed to enhance representational learning by providing new deep learning architectures or loss functions.

Self-supervised Depth Estimation: This method is based on the Structure from Motion (SFM) framework. In this approach, the problem is framed as learning to create a novel view from a sequence of images. The main task is to construct the target view I_t from the original view by taking frames in multiple time steps I_t-1, I_t+1 and employing a learned transformation from a pose deep learning neural network to perform the frame warping.

Self-supervise Depth Estimation: In this approach, the depth is estimated from the monocular camera based on the stereo information. Therefore, instead of using a sequence of frames as input, the model will estimate the disparities (left and right) d_l, d_r only using the left RGB frame, where a spatial transformer neural network warps the RGB frame pair I_l, I_r based on the disparity. This approach is based on the assumption that the baseline must be horizontal and known. The frame pair have to be rectified so that the transformation over the disparity is valid.

In this study, two methods are implemented to estimate the depth map from a single RGB-frame using an adaptive deep learning approach. This work depends on solving depth prediction as an image reconstruction problem. The main idea is that, we will try to learn how we could be able to obtain the right image from left, and when that is achieved, it means learning something about the 3D shape of the view is done. Therefore, the depth is obtained by inferring the disparities that match the left image to the right one and vice-versa.

A. Depth Estimation based on Semi-Supervised Learning

In this approach, a semi-supervised learning model is implemented to predict depth maps from a single frame without prior knowledge of the surroundings, by leveraging knowledge from both supervised and unsupervised learning. This approach is achieved by using a few annotated depth dataset and stereo pairs of RGB images provided by the KITTI dataset [5] as shown in Figure 3. The model is trained with LIDAR data which provides sparse depth information and the stereo image pair to predict high depth information during the inference phase.

In addition to the training, we try to explore the left-right consistency in a stereo reconstruction through a loss function. The evaluation of our model is being tested on the popular KITTI dataset which tests images and corresponding to the test images the depth maps are predicted.

B. Result & Evaluation based on Semi-Supervised Learning

For evaluating the performance of our model, 5 equation metrics have been used, namely RMSE, RMSE (log), accuracy, ARD and SRD. RMSE metrics calculate the number of information of depth which we produced through our model and compare it ground truth depth for the total number of pixels in that image. In the RMSE metrics, the

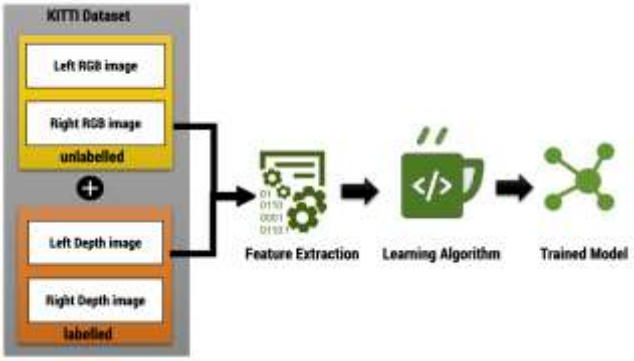

Figure 3: Semi-Supervised Training framework for Depth Estimation.

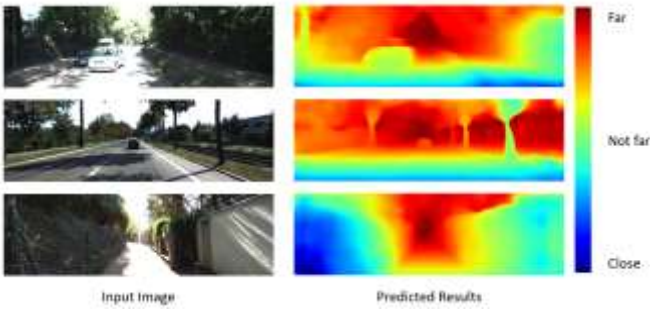

Figure 4: Semi-Supervised results for depth estimation.

Method	Lower is better				Higher is better		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen et al. Coarse	0.214	1.605	6.563	0.292	0.673	0.884	0.95
Eigen et al. Fine	0.203	1.548	6.307	0.282	0.702	0.89	0.95
Liu et al	0.201	1.584	6.471	0.273	0.68	0.898	0.96
Liana et al	0.179		4.46	0.077			
Garg et al	0.169	1.08	5.104	0.273	0.74	0.904	0.96
Godard et al	0.143	1.369	5.849	0.242	0.818	0.929	0.96
Godard et al + CS	0.136	1.532	5.763	0.239	0.816	0.935	0.96
Godard et al + CS + Post Processing	0.124	1.388	5.682	0.234	0.842	0.941	0.97
Kuznietsov et al	0.089	0.478	3.61	0.138	0.906	0.98	0.99
Amiri et al	**0.078**	**0.417**	**3.464**	**0.126**	**0.923**	**0.984**	**0.99**
CE Depth	0.083	0.601	3.558	0.136	0.915	0.973	0.99

Figure 5: KITTI Split evaluation results.

lower output value shows the model performance is good and when the value is high the model is not a good model. In the ARD and SRD which calculate the absolute and squared relative difference of the generated depth to the ground truth depth. In this scenario, the lower the value of ARD and SRD show a model is performing good and higher value shows that the model needs some improvement. The minimum amount of depth which our model can get is zero meter and the maximum depth is about 80 meter in a given scene. This evaluation is done base on the KITTI split [6].

The splits comprised of 28689 left depth image and Right depth image which is a supervised part of the model. For the unsupervised, we have to use the 28689 Left RGB image and Right RGB image. Figure 4 illustrates the input image on the left side and the predicted output depth map on the right side. The close (blue color) in this means zero meters from the viewpoint and the far (dark red color) means the 80-meter depth information from the viewpoint. Figure 5 shows us the accuracy of our model in comparison to the other researchers' work. According to the evaluation metrics, the RMSE value

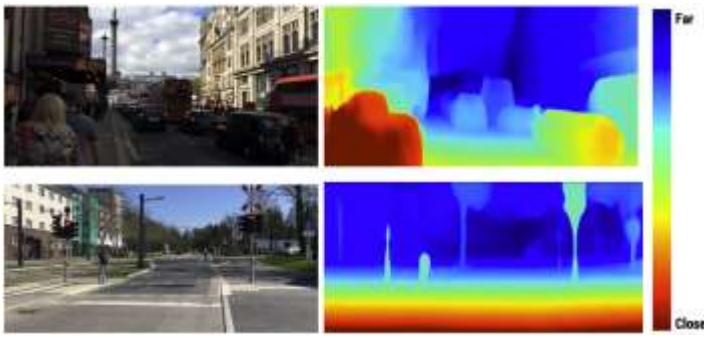

Figure 6: Semi-Supervised updated results.

Method	Lower is better				Higher is better		
	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou et. al (w/o PoseNet)	0.221	2.226	7.527	0.294	0.676	0.885	0.954
Zhou et. al	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Yin et. al	0.155	1.296	5.857	0.233	0.793	0.931	0.973
CE- Model 1	0.2189	3.4685	7.3869	0.2912	0.7245	0.8930	0.9516
CE- Model 2	0.1852	1.0923	5.007	0.2574	0.7257	0.9069	0.9678
CE- Model 3	0.1906	1.194	5.2243	0.2694	0.704	0.901	0.963
CE- Model 4	0.1809	1.1025	5.1747	0.2623	0.7211	0.9058	0.9654

Figure 7: Unsupervised Learning evaluation results.

was 3.928, and when compared to the other results, we are very close to the state of the art. The RMSE value defines how many per pixel in the image we are able to calculate the depth pixels.

V. CURRENT PROGRESS

A. Updated Results based on Semi-Supervised Learning

In current progress, we try to improve the prediction model by employing a bunch of tips and techniques such as data transform, feature selection, and tuning algorithm, and our current updated results are shown in Figure 6. Furthermore, we try to transform the model into a light version to run on Edge computing hardware such an NVIDIA Jetson Nano model and evaluate our solution with respect to resource consumption and run times.

B. Depth Estimation based on Unsupervised Learning

As it is illustrated previously, it is not sufficient to use the ground truth depth information and even it is hard and costly (such as LIDAR). However, we could be able to learn from unlabeled streaming video clips. As we observed, the individual frames within each streaming video clip are not arbitrary but are projections the same 3D sense from various angles that means, if we were able to model the 3D sense and camera perspectives of the sequences of video frames, we could correctly synthesize the video frames through geometric projection. This is known as geometry-based view synthesis and has been studied extensively in the literature. Our work follows a similar approach, as described in [7], but in a more flexible way. However, we could be able to train our model on videos captured by a single moving camera. This solution depends on the unsupervised learning approach.

C. Premilinary Result based on Unsupervised Learning

In this approach, different models for depth and pose networks are jointly trained from the unlabelled dataset, in order to produce a better depth estimation, in this solution no ground truth 3D or pose labels are available. Figure 7 shows us the accuracy of our model in comparison to the other researchers' work.

VI. CONCLUSION & FUTURE SCOPE

In the presented study, different approaches are studied and implemented to estimate the depth map from a single RGB-image using the adaptive deep learning network. This work depends on solving depth prediction as an image reconstruction problem. The self-supervision with stereo data is still required during the training time. However, this should be addressed and optimized in future work. Moreover, the transparent surfaces and occlusion regions will output invalid depths. These limitations and results should be improved in future work. In addition, we are going to develop pruning strategies that rely on embedded GPUs to minimize model complexity and consequently reduce the number of samples that are necessary for training to maximize the accuracy and speed of deep learning algorithms. Moreover, it is essential to estimate the full occupancy map of the scene and extend the presented approach to obtain the depth map from the video.

REFERENCES

[1] Saleh, S., Hadi, S., M., Amin Nazari, and Hardt, W. (2019). Outdoor Navigation for Visually Impaired based on Deep Learning. In: 6th International Conference Actual Problems of System and Software Engineering. [online] Moscow: IEEE, pp.397-406. Available at: http://ceur-ws.org/Vol-2514/.

[2] Shotton J. et al., "Efficient Human Pose Estimation from Single Depth Images," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 12, pp. 2821–2840, Dec. 2013.

[3] Weng, X. and Kitani, K. (2019) 'Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud', Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019. Institute of Electrical and Electronics Engineers Inc., pp. 857–866.

[4] Saleh, S., Khwandah, S., Heller, A., Mumtaz, A. and Hardt, W. (2019). Traffic Signs Recognition and Distance Estimation using a Monocular Camera. In: 6th International Conference Actual Problems of System and Software Engineering. [online] Moscow: IEEE, pp.407-418. Available at: http://ceur-ws.org/Vol-2514/.

[5] Menze M. and Geiger A., "Object scene flow for autonomous vehicles," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3061–3070.

[6] Jahani Amiri A., Shing Y. L., and Zhang H.. Semi-Supervised Monocular Depth Estimation with Left-Right Consistency Using Deep Neural Network.2019.

[7] Zhou T., Brown M., Snavely N., and Lowe D.G., "Unsupervised Learning of Depth and Ego-Motion from Video," UC Berkeley and Google, in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6612–6619.

[8] Saleh S. M., Khwandah S. A., Hardt W., Hilbrich M. and Lazaridis P. I., "Estimating the 2D Static Map Based on Moving Stereo Camera," 2018 24th International Conference on Automation and Computing (ICAC), Newcastle upon Tyne, United Kingdom, 2018, pp. 1-5, doi: 10.23919/IConAC.2018.8749004.