Monocular vision based 3D pose estimation for enhanced cyclist safety

Nikolina Crnogorac[1], Alexander Sing[1], Csaba Beleznai[1], Markus Vincze[2]

***Abstract*— Recent developments in Deep Learning demonstrate that accurately regressing 3D pose parameters from a monocular view is a feasible task. An estimated pose of specific objects with known dimensions from a mobile observer's viewpoint reveals relevant spatial relationship, contributing to an understanding of the surrounding environment. Therefore, monocular 3D pose estimation is an important enabler in safety-related task domains such as perception for autonomous driving and automated traffic monitoring. In this paper we present conceptual considerations, a baseline methodology, and results towards monocular vision based 3D pose estimation involving the safe interaction between cyclists and other vehicles. Furthermore, we propose an enhancement of cyclist detection via learning pose-annotated appearances from a dataset, where retro-reflective stripes mounted on the bicycle frame generate a spatially-extended visible pattern. This pattern is introduced to enhance detection recall and pose estimation accuracy under adverse visibility conditions such as low-light, fog and heavy rain.**

I. INTRODUCTION

Our proposed research endeavour relates to automated visual environment perception in complex and dynamic environments. Recent advances in image-based representation learning via Deep Learning are opening new perspectives towards reliably detecting, classifying and tracking multiple interacting objects under a wide range of observation conditions. The main potential in representation learning is given not merely by its accuracy, but perhaps even more by its inherent representational flexibility. This flexibility implies that multiple learning tasks (detection, tracking, segmentation) and a varying representational granularity (object type, pose, additional attributes) can be jointly formulated and directly inferred from image data.

In this study we present considerations, a research methodology and results on detecting the 3D pose of cyclists and other vehicles from a monocular camera view. In our setting, both bicycles and vehicles are equipped with a monocular camera setup and a mobile computing unit, in order to generate pose-attributed detection and tracking results which represent other nearby vehicles in a surrounding 3D spatial context (see Figure 1). To enhance the detection and pose estimation accuracy, we propose the generation of a large mixed cyclist dataset, containing pose-annotated real and synthetic images. This dataset shall also include images captured under adverse photometric conditions, where a specific retro-reflective pattern on the bicycle frame shall generate a visible structure encoding distance (via its dimensions) and pose (via perspective foreshortening), even at poor visibility conditions.

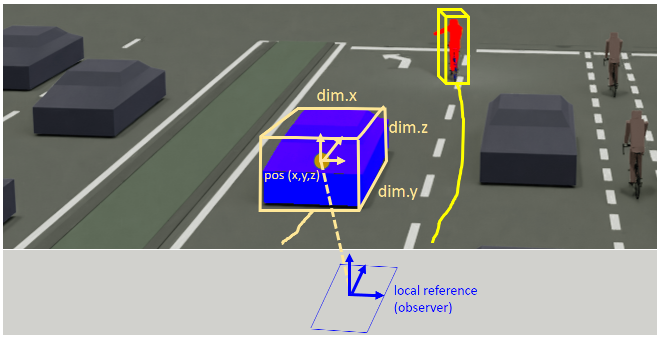

Fig. 1. Illustration depicting the monocular estimation of 3D pose from the viewpoint of a mobile observer.

3D Object detection based on a single perspective image (monocular image based detection) is considered to be a challenge. The object 3D pose is often represented by three spatial coordinates (x, y, z) and three orientations. In a street-level observation scenario, nevertheless, certain parameters can be assumed to be known: objects exist on the road surface $(z = 0)$, and only the yaw orientation (determining the direction of travel) is of relevance. However, even with a reduced set of parameters and known object dimensions, the object distance from the camera is a sensitive parameter to be estimated. Especially, as this distance (depth) uncertainty increases farther away from the camera. This monocular ambiguity can be lowered both by a diverse and accurately 3D-annotated dataset (via encoding precise spatial priors) and by algorithmic means, choosing parametric representations which can be more accurately regressed and mapped to a 3D world (birds-eye-view /BEV/) representation frame.

II. RELATED WORK

Monocular 3D pose estimation in street scenarios has emerged recently [3] and it is a subject of intense research. This surge of development partially stems from the emergence of pose-annotated datasets (initiated by the KITTI Vision Benchmark [1]), partially is due to the wider use of depth-sensing sensor modalities (stereo vision, LiDAR, Radar) which can be used to derive pose-annotations in an automated manner. The broad set of possible representation strategies and methodologies are well reviewed in [2]. A natural approach is the direct learning of the spatial transform

*This work was carried out within the Bike2CAV project, which is funded by the Austrian Ministry for Climate Action, Environment, Energy, Innovation and Technology (BMK) under the program “Mobility of the Future” and is managed by the Austrian Research Promotion Agency (FFG).

[1] Center for Vision, Automation & Control, AIT - Austrian Institute of Technology, Vienna, Austria, csaba.beleznai@ait.ac.at

[2] Automation and Control Institute, Vienna University of Technology, Vienna, Austria

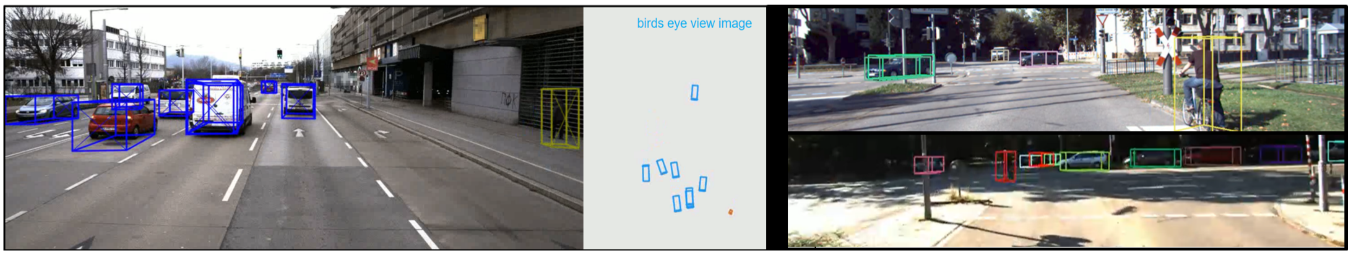

Fig. 2. Monocular pose estimation results without (left) and with a tracking step (right) shown for several test sequences.

from image to BEV space and hence removing problems associated with scale variation, such as in Pseudo-LIDAR [5] and OF-Transform [4]. Another simple representational objective is the direct regression of all pose parameters such as in Deep3DBox [3] and CenterNet [7].

Numerous open datasets have been proposed recently to support 3D pose estimation via learning. Table I provides an overview on the relevant datasets, especially highlighting the number of cyclist annotation instances.

III. METHODOLOGY

Pose-aware object detection and tracking: Based on a flexible convolutional encoder-decoder-type network [7], we formulate the learning objectives as a simultaneous object detection, pose parameter regression and appearance embedding task. The advantage in the simultaneous estimaton of these tasks is the shared representational backbone, leading to a representational homogeneity and run-time efficiency. Using the KITTI 3D dataset [1], we estimate the parameters of object location in the image, metric depth and the yaw angle α of object orientation, encoded as $(cos\alpha,\ sin\alpha)$. These attributes are learned for six object classes representing cyclists, pedestrians and diverse road vehicles. Flip augmentation is used to enrich the training dataset. Inspired by recent advances in target re-identification [6], we also incorporate a target association scheme exploiting an estimated low-dimensional appearance representation. Groundtruth tracking information from the training dataset is used to optimize this term by penalizing ID switches.

Low-light cyclist dataset: a key on-going research aspect is to explore simple ways, how to enhance cyclist detection and pose estimation accuracy under low-light conditions. To this end, we have devised a retro-reflective pattern design, which is low-cost, easy-to-deploy and exhibits a representational compatibility for learning. It means, that by enriching the dataset with pose-annotated reflective cyclist instances, the same learning framework can be used to create a pose-aware cyclist detector, with an extended range of illumination conditions. We also conceived an automated data acquisition procedure, which performs pose annotation of retro-reflective cyclist instances using a LiDAR sensor.

Dataset	**3D based on**	**3D annotations total / # of cyclist**	**Additional Information**
Cityscape3D (2020)	Stereo Vision	30k / 3960	
KITTI 3D (2012)	LIDAR	80k / 5400	Tracking information
nuScenes (2019)	LIDAR	1.4M / 7331	Tracking information
WaymoOpen (2019)	LIDAR	12.6M	
H3D (Honda) (2019)	LIDAR	1.1M / >10k	Tracking information
CADC (2020)	LIDAR, Camera	56k / 705	Winter conditions
RADIATE (2020)	LIDAR Stereo, Semiautomatic, CamShift	200k / 500	RADAR data (low spatial res.), diverse conditions, tracking information
ApolloScape (2018, 2019)	LIDAR	89k	
Agroverse (2019)	LIDAR	993k / 200	Tracking information
A*3D (2019)	Manual	230k / <100	Challenging conditions
Lyft L5 (2019)	LIDAR	1.3M	
A2D2 (2020)	LIDAR, Camera	12k / ≈700	
Astyx HiRes2019 (2019)	3D Camera, LIDAR, Stereo, Semi-automatic	5k	500 Synced Frames (Vis, Lidar, Radar), Custom RADAR

TABLE I

EXISTING 3D OPEN DATA-SETS FOR CYCLIST DETECTION

IV. RESULTS

Results are shown on our own and on the KITTI 3D [1] dataset exhibiting a street-level perspective (Fig. 2). Obtained detection results seem to capture the near-range context (vehicle constellations in the metric BEV space), which is relevant to estimate cyclist safety. However, with increasing range, estimated metric distances decay in spatial accuracy and temporal stability. Hence, an on-going research goal has been set to enhance representational capabilities refining and stabilizing spatial estimates.

V. SUMMARY AND OUTLOOK

In this paper we present results for pose-aware cyclist and vehicle detection and tracking, enabling an automated assessment of the local traffic context, and enhancing cyclist safety. Furthermore, we propose a low-light retro-reflective-enhanced cyclist dataset, which shall enhance pose-aware detectability under adverse lighting conditions.

REFERENCES

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[2] S.-h. Kim and Y. Hwang, "A survey on deep learning based methods and datasets for monocular 3d object detection," *Electronics*, vol. 10, no. 4, 2021.

[3] A. Mousavian, D. Anguelov, J. Flynn, and J. Košecká, "3d bounding box estimation using deep learning and geometry," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5632–5640.

[4] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3d object detection," *British Machine Vision Conference*, 2019.

[5] Y. Wang, W. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8437–8445.

[6] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *arXiv preprint arXiv:2004.01888*, 2020.

[7] X. Zhou, U. T. Austin, D. Wang, U. C. Berkeley, and U. T. Austin, "Object as Point," *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.