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.

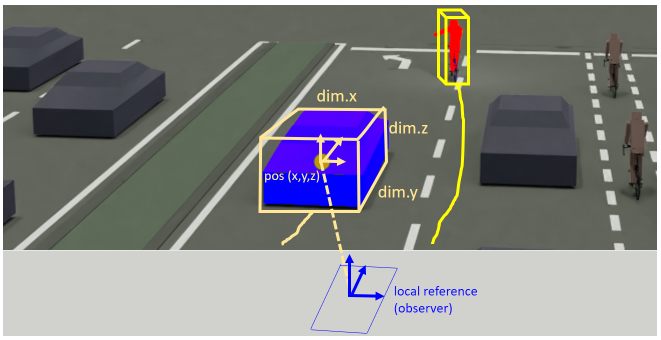

Fig. 1. Illustration depicting the monocular estimation of 3D pose from the viewpoint of a mobile observer.

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