2013 10th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)
October 31-November 2, 2013 / Ramada Plaza Jeju Hotel, Jeju, Korea

Resolving Scale Ambiguity for Monocular Visual Odometry

Sunglok Choi, Jaehyun Park, and Wonpil Yu

Intelligent Cognitive Technology Research Dept., ETRI, Republic of Korea
(E-mail: sunglok@etri.re.kr, Web: http://sites.google.com/site/sunglok)

Abstract - Scale ambiguity is an inherent problem in monocular visual odometry and SLAM. Our approach is based on common assumptions such that the ground is locally planar and its distance to a camera is constant. The assumptions are usually valid in mobile robots and vehicles moving in indoor and on-road environments. Based on the assumptions, the scale factors are derived by finding the ground in locally reconstructed 3D points. Previously, kernel density estimation with a Gaussian kernel was applied to detect the ground plane, but it generated biased scale factors. This paper proposes an asymmetric Gaussian kernel to estimate unknown scale factors accurately. The asymmetric kernel is inspired from a probabilistic modeling of inliers and outliers, that is, 3D point can comes from the ground and also other objects such as buildings and trees. We experimentally verified that our asymmetric kernel had almost twice higher accuracy than the previous Gaussian kernel. Our experiments was based on an open-source visual odometry and two kinds of public datasets.

Keywords - scale ambiguity, monocular visual odometry, monocular visual SLAM, asymmetric kernel

1. Introduction

Visual odometry has been popularly investigated because of its versatile applications to mobile robots and autonomous vehicles. It provides more accurate trajectory than other odometers such as wheel encoders and inertia measurement units (IMU). Visual odometry is implemented with two types of camera configurations: monocular and binocular (a.k.a. stereo). Compared to the monocular configuration, stereo cameras give more accurate trajectory with less computational burden. However, monocular visual odometry needs to be studied more because of several reasons. At first, many products and systems equipped with only a single camera, not a stereo camera (e.g. cellular phones and car black-box systems). Even though some have multiple cameras, they are non-overlapped so that they are not the stereo configuration. Moreover, the monocular configuration enables more compact size without additional lens and image sensor. Therefore, it can be easily embedded to other devices and products with more inexpensive cost.

However, monocular visual odometry inevitably suffers from scale ambiguity. The monocular configuration cannot identify the length of translational movement (a.k.a. scale factor) only from feature correspondences. Its example is presented in Figure 1. Therefore, in monocular visual odometry and SLAM, there have been numerous researches to solve the ambiguity. They mostly tackled the problem with additional information came from extra sensors or known motion/structures or constraints. Extra sensors fundamentally solved the ambiguity, but this approach can break one of the strongest advantages of the monocular configuration, compact size and less cost. Without extra sensors, a known initial scale factor can resolve the ambiguity, and its sequential feature mapping enables to estimate scale continuously. However, this approach suffers from another problem, scale drift, due to error accumulation in cumulative scale estimation. Additional constraints or assumptions sometimes help to overcome the scale ambiguity. An example of constraints is the locally planar ground assumption, which is mostly available in indoor and on-road environments. The previous researches are analyzed in Section 2 in detail.

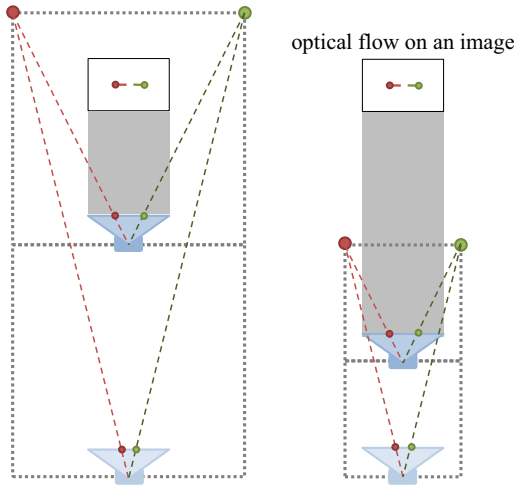


Fig. 1: Translation with two different scale factors exhibits the same feature correspondence. (a) The first camera move 2 meters forward in the 2-meter-wide tunnel. (b) The second camera move 1 meter forward in the 1-meter-wide tunnel.

This paper proposes an asymmetric Gaussian kernel to resolve the scale ambiguity. Without extra sensors and scale drift, our approach can estimate scale factors continuously with additional assumptions. We assume that the ground is locally flat so that a camera mounted on robots and vehicles is at constant distance above the ground. Under the assumptions, the scale factors are simply derived by finding the ground plane in locally reconstructed 3D map. The derivation is explained in Section 3.1. The previous Gaussian kernel generated biased scale factors, but our asymmetric kernel estimate scale factors

978-1-4799-1197-4/13/$31.00 ©2013 IEEE

accurately without the biased error. Our asymmetric kernel is based on probabilistic models of features from the ground and other objects such as buildings and trees so that it can reduce effect of features from other objects. Two kernels, Gaussian and asymmetric, are introduced in Section 3.2, respectively. Finally, Section 4 demonstrates effectiveness of our proposed asymmetric kernel. The experiments include evaluation of two kernels with two public datasets: the KITTI odometry dataset (on-road) and the ualberta-csc-flr3-vision dataset (indoor).

2. Related Works

A number of researches have investigated the scale ambiguity in monocular visual odometry and SLAM. Basically, the ambiguity is resolved by imposing additional information. We categorize the previous works according to types of additional information.

Extra Sensors Extra sensors can simply give scale factors if they can measure the scale directly or indirectly. For example, Scaramuzza et al. [1] utilized a speed meter for measuring the scale factors directly. A speed meter is intrinsically embedded in a vehicle and its measurement is available via CAN communication. IMUs such as accelerometers also provide scale factors through simple integration or sensor fusion with visual odometry. Additional sensors are against the benefit of a single camera, small and inexpensive, so they are preferred when they are inherently equipped in systems and platforms (e.g. IMUs in smart phones and speed meters in vehicles).

Known Initials It is also possible to estimate scale factors when its initial value is given or location of initially observed features is known. From the initial motion or structures, scale factors of the next frames are sequentially derived by filtering or optimization. For example, PTAM [2] assumes that translation between the first and second frames is 10 centimeters. From the initial motion, PTAM reconstructs a 3D feature map so that it is able to estimate correctly scaled motion from the map. Moreover, the map is evolved through sequential estimation of camera motion and features location. Similarly, MonoSLAM [3] starts from landmarks whose 3D position are known. From the known landmarks, MonoSLAM estimates correctly scaled motion and updates its feature map including unknown landmarks. PTAM and MonoSLAM are able to estimate correct scale factors only relied on the initially known motion and landmarks, but their sequential motion estimation can have drift error caused from error accumulation of scale factors. Therefore, both approaches are unsuitable for visual odometry usually working with few closed-loop paths. (e.g. Mars exploration).

Constraints Additional constraints and assumptions can also resolve scale ambiguity. One of common assumptions is that the height of a camera from the ground is almost constant. The assumption is valid for the camera mounted on indoor robots and on-road/rail vehicles. For example, Choi et al. [4] and Kitt et al. [5] adopted the assumption and applied it to planar homography of the ground. They also assumed that the ground surface is locally flat for planar homography.. To estimate planar homography, Choi et al. used correspondence of feature points and Kitt et al. utilized local patch matching. A well-known visual odometry library, LIBVISO2 [6], was also based on the assumption of constant camera height. Instead of planar homography, it found the ground plane in locally reconstructed 3D feature points and its height compared to its real value leads scale factors. More details are described in Section 3.1. A kinematic constraint of vehicles is also considered as a source of scale factors. Scaramuzza [7] investigated Ackerman's steering kinematics to estimate scale factors in turning situations. Scale estimation with additional constraints does not require extra sensors and does not suffer from scale drift, but it is only applicable when the constraints are valid and satisfied in the working environments.

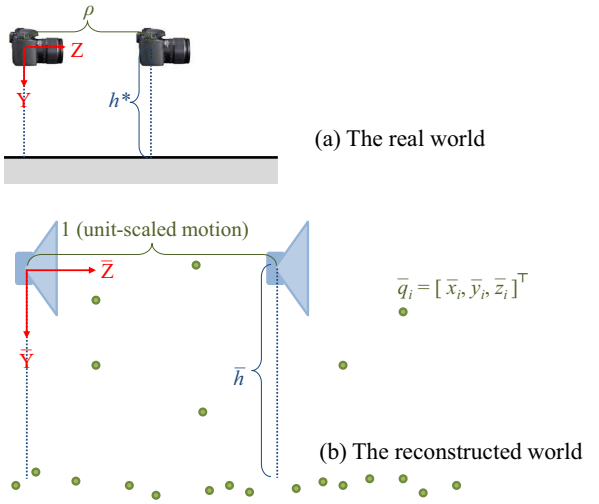


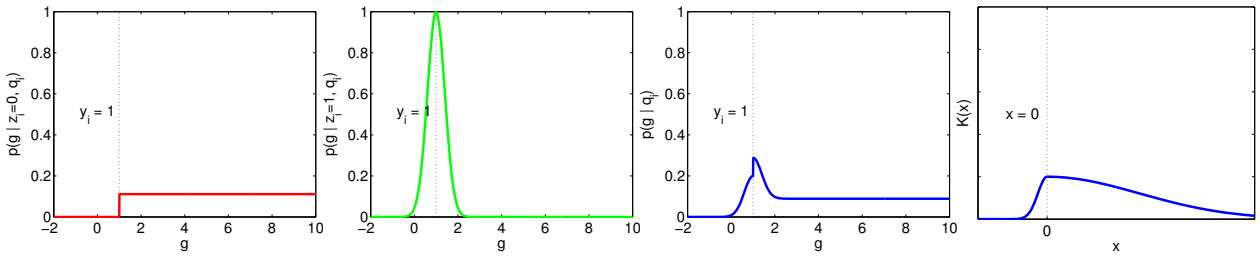
Fig. 2: The feature correspondences are reconstructed with unit-scaled motion. The real world (a) is proportional to the reconstructed world (b) in the amount of the scale factor, ρ.

3. Scale from Asymmetric Kernel

3.1 Scale from Camera Height

Our approach is based on two assumptions similar to the previous researches. First, the height of a camera from the ground is almost constant and written as h^*. Second, the ground is the most significant horizontal plane on almost every image frames. These two assumptions are usually valid for a camera mounted on vehicles in indoor and on-road environments.

Our scale estimation also utilizes a locally built feature map similar to LIBVISO2 [6]. Without correctly estimated scale factors, the local feature map is constructed from feature correspondences using triangulation and unit-scaled motion. In the feature map, i-th feature is written as $\bar{q}_i = [\bar{x}_i, \bar{y}_i, \bar{z}_i]^\top$. It is not same with its true location in the real world. The correctly scaled location is written as $p_i = \rho \bar{q}_i$ where ρ is the unknown scale



(a) $p(g|z_i = 0, \bar{q}_i)$ (b) $p(g|z_i = 1, \bar{q}_i)$ (c) $p(g|\bar{q}_i)$ (d) The asymmetric Gaussian kernel

Fig. 3: The overall likelihood (c) of the given $\bar{q}_i$ is the mixture of its inlier distribution (b) and outlier distribution (a) whose parameters are $\bar{y}_{max} = 10$, $\gamma_i = 0.2$, and $\sigma = 0.4$. The likelihood is approximated as the asymmetric Gaussian kernel (d).

factor which we want to know.

The unknown scale factor, ρ, is derived by comparing same structures in the real world and the reconstructed world. The real and reconstructed worlds are under the relationship of similarity as shown in Figure 2, and the ground is a good reference structure due to our assumptions. Especially, the height of a camera from the ground, h^*, is a good metric to lead the scale factor because it is constant and known. When $\bar{h}$ is its corresponding distance in the reconstructed feature map, the scale factor is the ratio of these two values as follows:

$$\rho = \frac{h^*}{\bar{h}}. \tag{1}$$

Therefore, finding $\bar{h}$ is the most important for estimating accurate scale factors.

3.2 Camera Height from Asymmetric Kernel

Gaussian Kernel LIBVISO2 [6] calculates the camera height, $\bar{h}$, using Gaussian-weighted voting. Since the ground is horizontal, $\bar{y}_i$ is significant to detect the ground, a flat XZ plane. Based on the second assumption, the ground is at the most densely clustered point on the y coordinate in the reconstructed world. LIBVISO2 finds the mode of feature density using kernel density estimation (KDE) with the Gaussian kernel as follows:

$$\bar{h} = \arg\max_h \sum_i \kappa(h - \bar{y}_i), \tag{2}$$

where κ is the kernel function. LIBVISO2 uses the Gaussian kernel function defined as

$$\kappa(x) = \exp\left(-\frac{x^2}{2\sigma_h^2}\right) \text{ such that } \sigma_h = \frac{1}{50}\operatorname*{med}_i \|\bar{q}_i\|_1, \tag{3}$$

where a function, med, finds the median of given values, and $\|\ \|_1$ is the L1-norm of given vector. Feature points can be extracted from the ground and also surrounding objects such as buildings, trees, and cars. A feature point is an inlier when it is from the ground, but it is an outlier when it comes from other objects. The label of i-th feature point is written as z_i, and its value is 1 when it is an inlier and 0 when it is an outlier. The Gaussian kernel provides reasonable scale factors with small amount of outliers, but it gives biased results with higher rate of outliers. Its examples are presented in Figure 5. Since KDE with the symmetric kernel does not distinguish inliers from outliers, it causes the biased scale factors.

Asymmetric Kernel For unbiased estimation, we propose a novel asymmetric kernel instead of the Gaussian kernel. The asymmetric kernel is inspired from the probability distribution of the ground when a feature point $\bar{q}_i$ is given. If a feature point comes from other objects, the ground may exist below the feature point because the ground is the lowest plane in the given environments. Therefore, without any prior, the probability of the ground is uniform as like

$$p(g|z_i = 0, \bar{q}_i) = \begin{cases} \frac{1}{\bar{y}_{max} - \bar{y}_i} & \text{if } \bar{y}_i < g < \bar{y}_{max} \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where g is a random variable to represent the location of the ground on the y coordinate and $\bar{y}_{max}$ is the upper-bound of $\bar{y}$. In contrast, if a feature point is from the ground, the ground may be around the point. It is usually modeled by Gaussian distribution as follows:

$$p(g|z_i = 1, \bar{q}_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(g - y_i)^2}{2\sigma^2}\right). \tag{5}$$

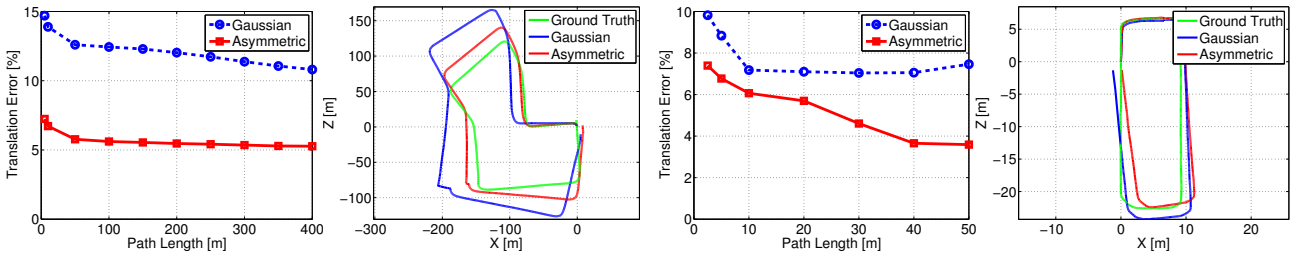
From the inlier and outlier distributions, the overall likelihood of the given $\bar{q}_i$ is the mixture of two distributions as like

$$p(g|\bar{q}_i) = (1 - \gamma_i)p(g|z_i = 0, \bar{q}_i) + \gamma_i p(g|z_i = 0, \bar{q}_i), \tag{6}$$

where γ_i is the prior probability of being an inlier, $p(g|z_i = 0, \bar{q}_i)$. Figure 3 presents two distributions and their overall likelihood. Finally, in the sake of simple computation, we approximate the complex likelihood (6) as an asymmetric Gaussian kernel as follows:

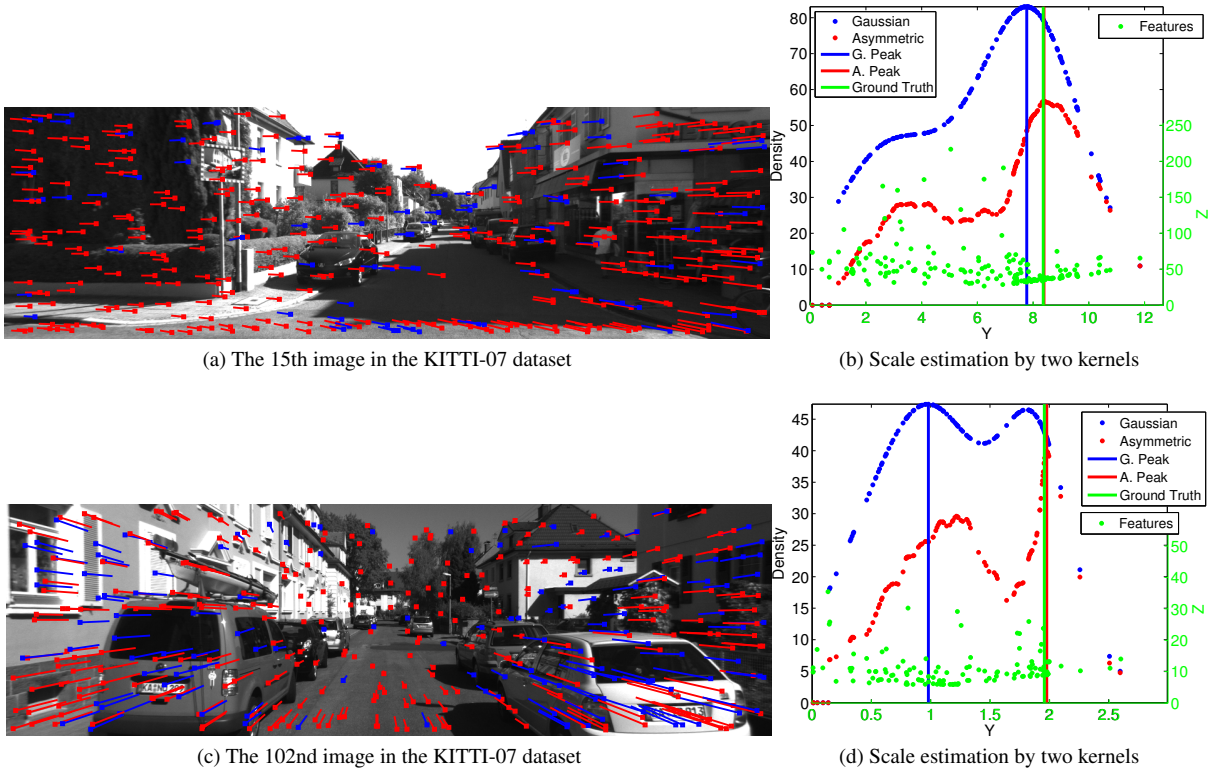
$$\kappa(x) = \begin{cases} \exp(-0.5x^2/\sigma_+^2) & \text{if } x > 0 \\ \exp(-0.5x^2/\sigma_-^2) & \text{otherwise} \end{cases}, \tag{7}$$

which is also presented in Figure 3d. The kernel width σ_+ should be bigger than σ_- to consider the probability of being outliers. We assign σ_+ as σ_h and σ_- as $0.01\sigma_h$. In contrast to the original likelihood, the approximated kernel only needs two parameters, σ_+ and σ_-, instead of $\bar{y}_{max}$, σ, and each prior probability of being inliers, γ_i.



(a) Trans. error in KITTI dataset (b) Trajectories in KITTI-07 dataset (c) Trans. error in ualberta dataset (d) Trajectories in ualberta dataset

Fig. 4: Translation errors in two datasets are presented in (a) and (c) with respect to various path lengths. Their estimated trajectories are also shown in (b) and (d) with their true trajectory. Since the KITTI odometry dataset consists of 11 sets of separate image sequences, we only present its 8th situation, KITTI-07.



(a) The 15th image in the KITTI-07 dataset (b) Scale estimation by two kernels

(c) The 102nd image in the KITTI-07 dataset (d) Scale estimation by two kernels

Fig. 5: Two examples of scale estimation, (b) and (d), are presented with its given images, (a) and (c). The first example contained the smaller number of outliers. In contrast, the second example had more outliers, and some of outliers from cars were horizontally aligned as like the ground. In Figure (a) and (c), feature correspondences are classified into correctly matched and wrongly matched features, respectively. In Figure (b) and (d), 3D feature points are presented in their reconstructed world using unit-scaled motion. Two feature densities by Gaussian and asymmetric kernels and their peaks are also presented in the locally reconstructed world.

4. Experiments

4.1 Configuration

We performed experiments to verify effectiveness of our asymmetric kernel in monocular visual odometry. LIBVISO2 [6], especially its monocular version, was adopted as a basic framework of visual odometry. We utilized its feature extraction/tracking and pose estimation without modification. In scale estimation, we additionally implemented our asymmetric kernel and compared it to the original Gaussian kernel. We used two public datasets: the KITTI odometry dataset [8] and ualberta-csc-flr3-vision dataset included in Radish [9]. The KITTI odometry dataset was acquired from a stereo camera mounted on a vehicle on roads, and we used only images from the left camera for monocular configuration. The dataset consists of 11 sets of image sequences with their true trajectories. Its total number of image frames is 23,201 and its travel distance is around 22,177 meters. The ualberta-csc-flr3-vision dataset was recorded from a single camera on an indoor environment, corridors. It has only one set of image sequences, and its number of frames is 513, and its travel distance is about 73 meters. Similar to the KITTI odometry benchmark [8], we calcu-

lated translational error of each piece of trajectories with respect to various path lengths. The translation error is the ratio of position error to the given path length so that is unit is percent. For example, if the translation error is 10 percent, its position error is 10 meters in 100-meter path length and 20 meters in 200-meter path length.

4.2 Results and Discussion

At first, our asymmetric kernel was almost 2 times more accurate than the Gaussian kernel in average. Translation errors of two kernels are described in Figure 4. We could observe that the Gaussian kernel mostly overestimated scale factors, which is also shown in two trajectories in Figure 4. The Gaussian kernel does not consider feature points from other objects and regards every feature as inliers. However, building, trees, and cars existed in almost all images and they generated feature points above the ground, that is, outliers. The outliers made KDE with the Gaussian kernel select $\bar{h}$ smaller than its truth. As shown in Equation 2, smaller $\bar{h}$ caused bigger ρ than its truth. It is also observed in examples of scale estimation in Figure 5. $\bar{h}$ was close to its truth with the small number of outliers, but it became seriously worse with higher rate of outliers.

Translation error in the ualberta-csc-flr3-vision dataset was slightly different from results in the KITTI odometry dataset. Basically, the difference resulted from variance of their given scale factors. The ualberta dataset was acquired every 1.5 meters, but the KITTI dataset was recorded every 0.1 seconds so that its scale factors varied from 0 meter up to 25 meters. In scale estimation, short scale factors are more difficult to estimate because the short scale factors lead inaccurate reconstructed feature maps. Moreover, since the KITTI dataset had long traversal with many image frames, its results were smoother than results of the ualberta dataset.

5. Conclusion

In this paper, we introduced the asymmetric Gaussian kernel for unbiased scale estimation. We also verified that our asymmetric kernel is more effective than the original Gaussian kernel in the view of accuracy. From two public datasets, scale estimation with our asymmetric kernel was almost twice more accurate than the Gaussian kernel.

As further works, we can improve the current approach by applying probabilistic models without approximation appeared in Figure 3. For more accurate pose and scale estimation, it is also possible to merge all available information together. For example, we can utilize unit-scaled motion from 2D-2D correspondence, our assumptions on the ground, and also correctly scaled motion from 2D-3D correspondence from known 3D points.

Acknowledgement

This work was supported partly by the R&D program of the Korea Ministry of Trade, Industry and Energy (MOTIE) and the Korea Institute for Advancement of Technology (KIAT). (Project: 3D Perception and Robot Navigation Technology for Unstructured Environments, M002300090) The author also would like to thank Andreas Geiger for sharing his LIBVISO2 and KITTI Vision Benchmark Suites and Jonathan Klippenstein for sharing the ualberta-csc-flr3-vision dataset.

References

[1] D. Scaramuzza, F. Fraundorfer, and R. Siegwart, "Real-time monocular visual odometry for on-road vehicles with 1-point RANSAC," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

[2] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *Proceedings of IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2007.

[3] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera slam," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 1–16, 2007.

[4] S. Choi, J. H. Joung, W. Yu, and J.-I. Cho, "What does ground tell us? monocular visual odometry under planar motion constraint," in *Proceedings of the International Conference on Control, Automation, and Systems (ICCAS)*, 2011.

[5] B. Kitt, J. Rehder, A. Chambers, M. Schonbein, H. Lategahn, and S. Singh, "Monocular visual odometry using a planar road model to solve scale ambiguity," in *Proceedings of European Conference on Mobile Robots (ECMR)*, 2011.

[6] A. Geiger, J. Ziegler, and C. Stiller, "StereoScan: Dense 3D reconstruction in real-time," in *Proceedings of IEEE Intelligent Vehicles Symposium (IV)*, 2011.

[7] D. Scaramuzza, "1-point-ransac structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints," *International Journal of Computer Vision (IJCV)*, vol. 95, pp. 74–85, 2011.

[8] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI Vision Benchmark Suite," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[9] A. Howard and N. Roy, "The Robotics Data Set Repository (Radish)," 2003. [Online]. Available: http://radish.sourceforge.net/

[10] H.-S. Kang, S.-W. Lee, and H. G. Hosseini, "Probability constrained search range determination for fast motion estimation," *ETRI Journal*, vol. 34, no. 3, 2012.