Localization in Autonomous Vehicles Using a Generalized Inner Product

Samuel Todd Flanagan, Drupad K. Khublani,
Jean-Francois Chamberland, Siddharth Agarwal, Ankit Vora

Abstract—**Fine localization in autonomous driving platforms is a task of broad interest, receiving much attention in recent years. Some localization algorithms use the Euclidean distance as a similarity measure between the local image acquired by a camera and a global map, which acts as side information. The global map is typically expressed in terms of the coordinate system of the road plane. Yet, a road image captured by a camera is subject to distortion in that nearby features on the road have much larger footprints on the focal plane of the camera compared with those of equally-sized features that lie farther ahead of the vehicle. Using commodity computational tools, it is straightforward to execute a transformation and, thereby, bring the distorted image into the frame of reference of the global map. However, this non-linear transformation results in unequal noise amplification. The noise profile induced by this transformation should be accounted for when trying to match an acquired image to a global map, with more reliable regions being given more weight in the process. This physical reality presents an algorithmic opportunity to improve existing localization algorithms, especially in harsh conditions. This article reviews the physics of road feature acquisition through a camera, and it proposes an improved matching method rooted in statistical analysis. Findings are supported by numerical simulations.**

I. INTRODUCTION

Autonomous vehicles have received considerable attention in recent years, with several companies deploying prototypes on public roadways. A common task for various autonomous platforms is fine localization at the scale of centimeters. Localization is the process by which an autonomous vehicle uses sensor data to determine its position within a map. Mathematically, this can be accomplished by calculating the inner product of sensor data and candidate sections of the map. The candidate section that produces the maximum or minimum inner product, depending on the formulation, is the most likely match. Inner products have been used in the past for matching in both localization and image registration, e.g., [1], [2].

Much of the existing results in this area focus on simultaneous localization and mapping (SLAM). Many implementations of SLAM have been developed over the past two decades. In 2001, Dissanayake et al. proposed a solution to the SLAM problem using an estimation-theoretic or Kalman filter based approach [3]. Since then, implementations such as DP-SLAM [4], Atlas [5], and Graph SLAM [6] have been

proposed. A current industry standard for SLAM, described in [7], is a version of Graph SLAM that achieves "reliable real-time localization with accuracy in the 10-cm range." In a subsequent article, the authors propose improvements such as using probabilistic maps to represent the environment that "increased robustness to environmental changes and dynamic obstacles," while retaining similar accuracy [8].

These autonomous vehicle systems use high-quality sensor arrays in favorable conditions to provide near-noiseless data for localization. However, as autonomous vehicles move closer to production, lower quality sensors will likely be utilized to reduce cost. Inexpensive cameras have been used in SLAM research (visual SLAM), yet proposed solutions struggle in challenging conditions [9]. In this paper, we use signal processing techniques to develop a novel algorithm that leverages a generalized inner product for localization. The proposed scheme outperforms the standard inner product for matching, with significant gains under adverse conditions.

II. PROBLEM FORMULATION

This work focuses on localization using a single camera with a predetermined global map that accurately represents a top-down view of the environment. Local images are captured in real-time, transformed, scaled, and matched with a section of the map. Herein, we treat gray-scale images, yet the proposed techniques can be extended to color images, image gradients, and processed images with minor changes to the implementation.

Our model has a focal plane and physical coordinate system with 2 and 3 dimensions, respectively. We use x, y, and z for the physical coordinate system and $\tilde{x}$ and $\tilde{y}$ for the focal plane of the camera. The origin of the physical coordinate system is centered on the pinhole of the camera, height h above a planar road as shown in Fig. 1. Parameter θ denotes the angle between the z-axis and the horizon.
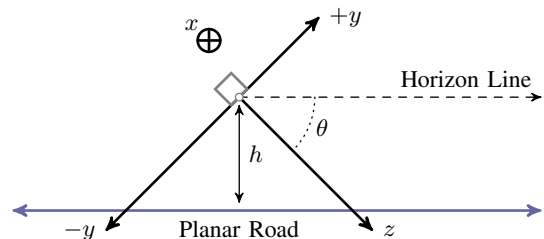

Fig. 1. This diagram illustrates the physical coordinate system and the camera orientation.

In our model, we use a pinhole camera with focal length, f, the distance between the pinhole and the focal plane. It is worth noting that, for this initial treatment, we disregard barrel and pincushion distortion in our analysis. The origin of the focal plane coordinate system is located at the center of the focal plane, $(0, 0, -f)$ in physical coordinates. Figure 2 shows the orientation of the coordinate systems. $\tilde{x}$ and $\tilde{y}$ are antiparallel to x and y which allows us to effectively ignore image inversion in the following sections.
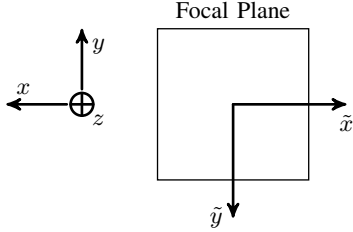

Fig. 2. The physical coordinate system is selected to align, partly, with the coordinate system of the focal plane, as shown above.

A perspective transformation models how objects appear when viewed from different positions [10]. Figure 3 highlights several aspects of the perspective transformation at hand. Nearby objects appear larger than those farther away. Lines parallel to the view plane remain parallel. Lines not parallel to the view plane are distorted.

Fig. 3. This image captures the effects of a camera perspective.

For our model, the perspective transformation that describes how objects are projected onto the focal plane, is given by

$$\tilde{x} = fx/z \qquad \qquad \tilde{y} = fy/z. \qquad (1)$$

III. FOCAL PLANE GEOMETRY

In this section, we develop a relationship between the area of a section of road and the footprint of its image on the focal plane. We describe the road surface shown in Fig. 1 with free variables x and z. Consequently, y can be expressed as $z \tan \theta - h \sec \theta$, where θ is the angle between the z axis and the horizon. We use this relation to rewrite part of (1) as

$$\tilde{y} = f \tan \theta - (fh/z) \sec \theta. \qquad (2)$$

We also need to express z as a function of $\tilde{y}$ for our discussion in Section IV. This relation is readily found to be

$$z = \frac{fh}{f \sin \theta - \tilde{y} \cos \theta}. \qquad (3)$$
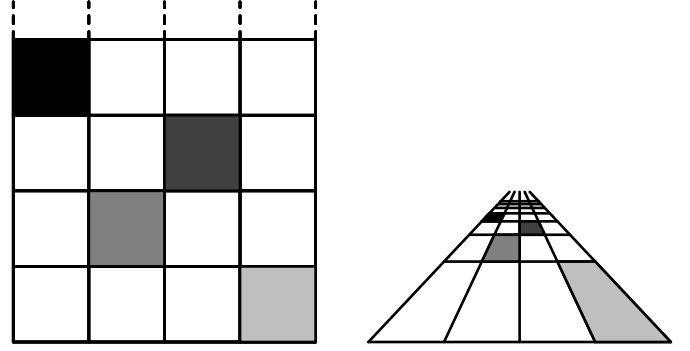

Fig. 4. The diagram on the left showcases an arbitrary road grid placed in front of an autonomous vehicle. The area of every square is kept constant. On the right, the drawing illustrates how the same grid pattern appears on the focal plane of the camera. While all the squares share the same physical area, their footprints on the focal plane differ.

Figure 4 depicts an arbitrary grid pattern that lies in front of the autonomous vehicle. It also shows the same grid pattern as acquired by the focal plane of the camera. This type of image distortion should be familiar to the reader because an analogous process governs human vision. It is especially pertinent to note that squares of equal area in the physical world can have vastly different footprints on the focal plane of the camera. In particular, squares that are farther away correspond to much smaller regions on the focal plane. This phenomenon has repercussions both in terms of signal acquisition and noise corruption.

In reference to calculus, the absolute value of the Jacobian determinant at point (x_1, z_1) expresses how much the area near (x_1, z_1) expands or contracts when projected onto the focal plane. The Jacobian of the perspective transformation, governed by (1) and (2), is given by

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \tilde{x}}{\partial x} & \frac{\partial \tilde{y}}{\partial x} \\ \frac{\partial \tilde{x}}{\partial z} & \frac{\partial \tilde{y}}{\partial z} \end{bmatrix} = \begin{bmatrix} \frac{f}{z} & 0 \\ -\frac{fx}{z^2} & -\frac{fh}{z^2} \sec \theta \end{bmatrix}. \qquad (4)$$

The Jacobian determinant, $\det(\mathbf{J})$, is found to be

$$\det(\mathbf{J}) = -f^2 h \sec \theta / z^3. \qquad (5)$$

Using (5) we compute the relationship between the area of a rectangular region $\mathcal{R} = [x_l, x_u] \times [z_l, z_u]$, which lies on the road in front of the autonomous vehicle, and the area of its projection on the focal plane as

$$\tilde{\mathcal{A}} = f^2 h \sec \theta \frac{(x_u - x_l)}{2} \left(\frac{1}{z_l^2} - \frac{1}{z_u^2} \right). \qquad (6)$$

Having developed this relationship, we can characterize the noise pattern associated with the perspective transformation.

IV. IMAGE ACQUISITION AND NOISE

The purpose of this section is to describe how the transformation from the physical world to the focal plane of the camera impacts the quality of the image. Suppose that, at a given time, features on the road surface are captured by $g(x, z)$. Then, at the same instant, the projection of the road

onto the focal plane of the camera can be expressed, using (1) and (3), as

$$\tilde{g}(\tilde{x}, \tilde{y}) = g\left(\frac{h\tilde{x}}{f\sin\theta - \tilde{y}\cos\theta}, \frac{h\tilde{y}}{f\sin\theta - \tilde{y}\cos\theta}\right).$$

The signal captured on the focal plane of the camera as a function of location $(\tilde{x}, \tilde{y})$ is given by $\tilde{g}(\tilde{x}, \tilde{y}) + N(\tilde{x}, \tilde{y})$, where N is two-dimensional white noise with power spectral density N_0. Consequently, the aggregate signal over a region becomes

$$S = \underbrace{\iint_{\tilde{R}} \tilde{g}(\tilde{x}, \tilde{y})d\tilde{x}d\tilde{y}}_{\text{amplitude}} + \underbrace{\iint_{\tilde{R}} N(\tilde{x}, \tilde{y})d\tilde{x}d\tilde{y}}_{\text{noise}}.$$

When the CCD sensor is operating within its linear region, as opposed to saturation, thermal noise is well-modeled as an additive zero-mean Gaussian random variable with variance $\sigma^2 = \tilde{A}N_0$. The 2D power spectral density parameter N_0 implicitly depends on a number of factors, yet it remains constant over the focal plane of the camera. To gain further intuition into the impact of the perspective transformation over performance, we examine a simplified model where the spatial signal has a constant magnitude in the physical world.

A. Constant Amplitude Regions

Consider a situation where a rectangular area on the planar road possesses a constant magnitude. Assume this rectangle corresponds to region $\mathcal{R} = [x_1, x_u] \times [z_1, z_u]$. That is, $g(x, z) = a$ for $(x, z) \in \mathcal{R}$. We wish to compute the effective signal-to-noise ratio (SNR) corresponding to this area as observed by the focal plane. We begin by computing the energy of the signal component, which is equal to

$$\left(\iint_{\tilde{\mathcal{R}}} \tilde{g}(\tilde{x}, \tilde{y})d\tilde{x}d\tilde{y}\right)^2 = \left(\iint_{\tilde{\mathcal{R}}} a d\tilde{x}d\tilde{y}\right)^2 = a^2\tilde{A}^2. \quad (7)$$

An explicit expression for $\tilde{A}$ appears in (6). By assumption, the noise component is independent of $\tilde{g}(\tilde{x}, \tilde{y})$. As mentioned above, the noise is zero-mean with variance $\tilde{A}N_0$. The effective signal-to-noise ratio for the component of the image corresponding to region $\mathcal{R}$ is then given by

$$\text{SNR} = \frac{a^2\tilde{A}}{N_0} = \frac{a^2}{N_0}\frac{f^2h}{\cos\theta}\frac{(x_u - x_1)}{2}\left(\frac{1}{z_1^2} - \frac{1}{z_u^2}\right). \quad (8)$$

The main insight is that the SNR decreases dramatically as a function of z. When matching a local image to a global map, this inherent phenomenon should be taken into consideration. This is especially true for situations where average SNR is low, such as poor light conditions or images acquired by inexpensive cameras. To further illustrate this point, we explore the structure of an optimal decision rule for tesselated images with piecewise constant amplitude.

B. Maximal Ratio Combining

Having developed a suitable characterization of the SNR for a uniform region, we turn to the scenario where the road plane is partitioned into multiple rectangular regions in a manner akin to Fig. 4. To facilitate analysis, we calibrate observations to be centered around zero. We assume every

square is randomly assigned a constant amplitude value independent of other regions. Because the road plane is acquired by a pinhole camera, different square regions feature different SNRs, as described above. The ensuing localization process entails matching the acquired image to a quantized global map. Under current assumptions, the localization task becomes a canonical vector Gaussian classification problem.

The tesselated road portion acquired by the camera can be reshaped into a vector, which we denote by $\mathbf{a}$. This gives rise to a noisy observation vector

$$\mathbf{v} = \mathbf{a} + \mathbf{n},$$

where $\mathbf{a}$ is the amplitude vector and $\mathbf{n}$ denotes an additive Gaussian noise vector. In this equation, the value of indexed element a_k in $\mathbf{a} = (a_1, \ldots, a_m)$ corresponds to the amplitude of a rectangular region on the road surface. Component n_k, which accounts for the effects of the noise process associated with the CCD sensor over the footprint of rectangular region $\tilde{\mathcal{R}}_k$, has mean zero and variance

$$\sigma_k^2 = N_0/\tilde{A}_k.$$

Since the tesselated regions do not overlap on the focal plane, the components of noise vector $\mathbf{n}$ are independent from one another. The effective SNR for region $\mathcal{R}_k$ is then given by

$$\text{SNR}_k = a_k^2\tilde{A}_k/N_0$$

which is equivalent to (8).

In estimating the location of the vehicle, we assume that candidate locations are defined at the level of the road tesselation, or a constant integer multiple thereof. This way, candidate vectors all share a similar structure with regions of constant amplitude. The likelihood for candidate grid location $\hat{\mathbf{u}}$ is

$$\mathcal{L}(\hat{\mathbf{u}}|\mathbf{v}) = \frac{1}{\sqrt{2\pi\prod_k\sigma_k^2}}\exp\left(-\frac{1}{2}\sum_k\frac{(\mathbf{v}_k - \hat{\mathbf{u}}_k)^2}{\sigma_k^2}\right)$$

$$= \frac{1}{\sqrt{2\pi\prod_k\sigma_k^2}}\exp\left(-\frac{f^2h\sec\theta}{2N_0}\sum_k\mathbf{G}_{k,k}(\mathbf{v}_k - \hat{\mathbf{u}}_k)^2\right)$$

where we have implicitly defined elements

$$\mathbf{G}_{k,k} = \frac{(x_{u,k} - x_{1,k})}{2}\left(\frac{1}{z_{1,k}^2} - \frac{1}{z_{u,k}^2}\right) = \frac{\tilde{A}_k}{f^2h\sec\theta}. \quad (9)$$

The maximum likelihood (ML) decision rule for this classification task can be expressed as

$$\hat{\mathbf{u}}_{\text{ML}}(\mathbf{v}) = \arg\min_{\hat{\mathbf{u}}}\|\mathbf{v} - \hat{\mathbf{u}}\|_{\mathbf{G}}. \quad (10)$$

In the representation above, $\|\cdot\|_{\mathbf{G}}$ is the norm induced by the generalized inner product $\langle\mathbf{w}_1|\mathbf{w}_2\rangle_{\mathbf{G}} = \mathbf{w}_2^{\mathrm{T}}\mathbf{G}\mathbf{w}_1$, where Gramian matrix $\mathbf{G}$ is a positive-definite diagonal matrix whose non-zero entries are given by (9).

The key insight revealed through this analysis is that a weighted inner product and its induced norm $\|\cdot\|_{\mathbf{G}}$ should be used in the localization process rather than the standard inner product and the Euclidean norm. The weights of the generalized inner product are dictated by the physics of the camera and can accommodate fine or coarse granularity. At this point, it is appropriate to assess the potential gains associated with this algorithmic improvement.

C. Preliminary Performance Assessment

In this section, we assess performance for the model presented above. We assume the rectangular regions have constant magnitude, i.e., $|a_k| = a$ for all locations. From (10), we gather that the true location is selected by the ML decision rule whenever

$$\|\mathbf{v} - \mathbf{u}^*\|_{\mathbf{G}}^2 \leq \|\mathbf{v} - \hat{\mathbf{u}}\|_{\mathbf{G}}^2 \quad \forall \hat{\mathbf{u}} \neq \mathbf{u}^*. \tag{11}$$

If we look at one alternate location at a time, the condition of (11) reduces to $\|\mathbf{v} - \mathbf{u}^*\|_{\mathbf{G}}^2 - \|\mathbf{v} - \hat{\mathbf{u}}\|_{\mathbf{G}}^2 < 0$, or alternatively,

$$0 < \langle \mathbf{u}^* - \hat{\mathbf{u}} | \mathbf{u}^* \rangle_{\mathbf{G}} + \langle \mathbf{u}^* - \hat{\mathbf{u}} | \mathbf{n} \rangle_{\mathbf{G}}.$$

The first component is a known constant, whereas the second component is a zero-mean Gaussian random variable. The noise component can be rewritten as $\sum_k (u_k^* - \hat{u}_k)\mathbf{G}_{k,k} n_k$ and its variance becomes

$$\sum_k (u_k^* - \hat{u}_k)^2 \mathbf{G}_{k,k}^2 \sigma_k^2 = \frac{N_0}{f^2 h \sec \theta} \sum_k (u_k^* - \hat{u}_k)^2 \mathbf{G}_{k,k}.$$

We can then express the probability of error given a single alternative as

$$1 - \mathbf{\Phi}\left(\sqrt{\frac{f^2 h \sec \theta}{N_0}} \frac{\langle \mathbf{u}^* - \hat{\mathbf{u}} | \mathbf{u}^* \rangle_{\mathbf{G}}}{\|\mathbf{u}^* - \hat{\mathbf{u}}\|_{\mathbf{G}}} \right). \tag{12}$$

In comparison, if a standard inner product is employed as the basis for classification, the probability of error becomes

$$1 - \mathbf{\Phi}\left(\sqrt{\frac{f^2 h \sec \theta}{N_0}} \frac{\langle \mathbf{u}^* - \hat{\mathbf{u}} | \mathbf{u}^* \rangle}{\|\mathbf{u}^* - \hat{\mathbf{u}}\|_{\mathbf{G}^{-1}}} \right). \tag{13}$$

D. Simulated Performance

We compare the performance of the standard and generalized inner product in Fig. 6. Performance is defined by the probability of error as described in Section IV-C. We use the following parameters for observations of 66 squares.

- h = 58.3095 cm
- θ = 35.9020°
- Field of View: 39.2962° (vertical), 70.5288° (horizontal)
- Focal Length = 0.0367 cm
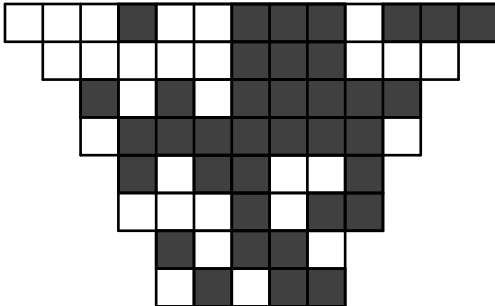- Square Side Length = 20 cm
- N_0 = 0.0018

Fig. 5. This grid offers a sample observation of whole squares acquired by the camera (after rectification). Note the similarity with the photo in Fig. 7.

We randomly generate $\mathbf{u}^*$ and $\hat{\mathbf{u}}$ vectors with amplitudes $\pm a$ and calculate the probabilities of error with (12) and (13). We perform 10,000 samples for every magnitude, with a ranging from 0.1 to 10.
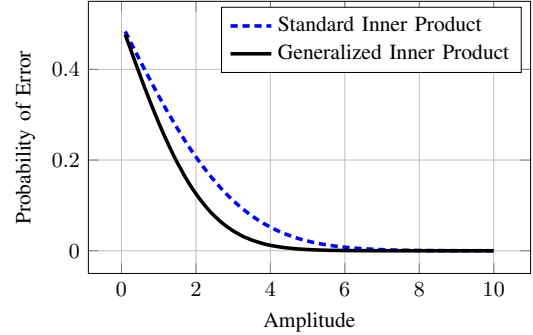

Fig. 6. This figure shows the generalized inner product outperforming the standard inner product. SNR increases with amplitude squared.

E. Experimental Results

Experimental results are obtained using a painted global map and images captured with a camera configuration similar to the parameters listed in Section IV-D. Preliminary results are aligned with the findings from our simulations.
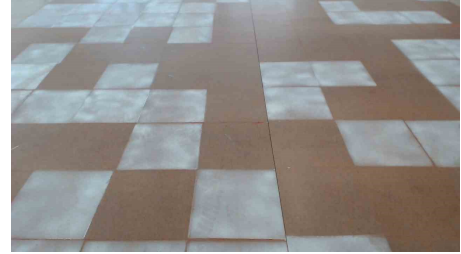

Fig. 7. This picture is a sample image of the global map, with its characteristic square sections, taken for our analysis.

V. CONCLUSIONS

In this article, we discussed the physics of the perspective transformation as it pertains to the acquisition of road images. We characterized the relationship between the area of a rectangular section of road and its footprint on the focal plane of the camera. This mapping underlies the phenomenon whereby objects farther from the camera appear smaller than nearer equally-sized objects. It also affects the signal quality of different sections of the road. This, in turn, creates an opportunity for an algorithmic improvement for localization tasks that are based on pattern matching. In particular, we proposed a generalized inner product that takes advantage of this nonuniform signal quality. This novel generalized inner product is optimal under certain conditions and it outperforms existing methods for localization in noisy conditions. Our findings are supported by numerical simulations. The proposed algorithmic framework and its potential improvements are especially significant in harsh environments such as low-light conditions and adverse weather situations, where future autonomous vehicles are poised to operate.

REFERENCES

[1] Stefan Krüger and Andrew Calway, "Image registration using multiresolution frequency domain correlation.," in *BMVC*, 1998, pp. 1–10.

[2] Kurt Konolige and Ken Chou, "Markov localization using correlation," in *IJCAI*, 1999, vol. 99, pp. 1154–1159.

[3] M. W. M. Gamini Dissanayake, Paul Newman, Steven Clark, Hugh F. Durrant-Whyte, and M. Csorba, "A solution to the simultaneous localization and map building (slam) problem," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 3, pp. 229–241, 2001.

[4] Austin Eliazar and Ronald Parr, "Dp-slam: Fast, robust simultaneous localization and mapping without predetermined landmarks," in *IJCAI*. Acapulco, Mexico, 2003, vol. 3, pp. 1135–1142.

[5] Michael Bosse, Paul Newman, John Leonard, and Seth Teller, "Simultaneous localization and map building in large-scale cyclic environments using the atlas framework," *The International Journal of Robotics Research*, vol. 23, no. 12, pp. 1113–1139, 2004.

[6] Sebastian Thrun and Michael Montemerlo, "The graph slam algorithm with applications to large-scale mapping of urban structures," *The International Journal of Robotics Research*, vol. 25, no. 5-6, pp. 403–429, 2006.

[7] Jesse Levinson, Michael Montemerlo, and Sebastian Thrun, "Map-based precision vehicle localization in urban environments," in *Robotics: Science and Systems*. Citeseer, 2007, vol. 4, p. 1.

[8] Jesse Levinson and Sebastian Thrun, "Robust vehicle localization in urban environments using probabilistic maps," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 4372–4378.

[9] Jorge Fuentes-Pacheco, José Ruiz-Ascencio, and Juan Manuel Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 55–81, 2015.

[10] D. Hearn, M. P. Baker, and W. R. Carithers, *Computer Graphics with OpenGL*, chapter Perspective Projections, pp. 327–340, Prentice Hall, 2011.

[11] Kichun Jo, Yongwoo Jo, Jae Kyu Suhr, Ho Gi Jung, and Myoungho Sunwoo, "Precise localization of an autonomous car based on probabilistic noise models of road surface marker features using multiple cameras," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 6, pp. 3377–3392, 2015.

[12] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*. Citeseer, 2011, vol. 11, p. 2.

[13] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[14] Jakob Engel, Thomas Schöps, and Daniel Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European Conference on Computer Vision*. Springer, 2014, pp. 834–849.

[15] Georg Klein and David Murray, "Parallel tracking and mapping for small ar workspaces," in *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. IEEE Computer Society, 2007, pp. 1–10.