TOWARDS AN ACCURATE LOW-COST STEREO-BASED NAVIGATION OF UNMANNED PLATFORMS IN GNSS-DENIED AREAS

Z. Shtain*, S. Filin

Mapping and Geo-Information Engineering, Technion – Israel Institute of Technology, Haifa, Israel
(zachis, filin)@technion.ac.il

Commission II

KEY WORDS: stereo, drones, localization, bundle adjustment, accuracy, simultaneous localization and mapping

ABSTRACT:

While lightweight stereo vision sensors provide detailed and high-resolution information that allows robust and accurate localization, the computation demands required for such process is doubled compared to monocular sensors. In this paper, an alternative model for pose estimation of stereo sensors is introduced which provides an efficient and precise framework for investigating system configurations and maximize pose accuracies. Using the proposed formulation, we examine the parameters that affect accurate pose estimation and their magnitudes and show that for standard operational altitudes of $\sim$50 m, a five-fold improvement in localization is reached, from $\sim$0.4–0.5 m with a single sensor to less than 0.1 m by taking advantage of the extended field of view from both cameras. Furthermore, such improvement is reached using cameras with reduced sensor size which are more affordable. Hence, a dual-camera setup improves not only the pose estimation but also enables to use smaller sensors and reduce the overall system cost. Our analysis shows that even a slight modification in camera directions improves the positional accuracy further and yield attitude angle as accurate as $\pm$6' (compared to $\pm$20'). The proposed pose estimation method relieves computational demands of traditional bundle adjustment processes and is easily integrated with other inertial sensors.

1. INTRODUCTION

The capabilities and availability of small unmanned aircraft and platforms have seen a dramatic rise in recent years, with the quadcopters becoming an everyday mapping utility for professionals and amateurs alike (Barry et al., 2015). Lightweight cameras become the preferable payload that provides detailed and high-resolution information which is used for a vast range of applications including, city modeling, risk zone assessment, archeology, cultural heritage, etc. (Gerke et al., 2016).

Platform navigation often relies on GNSS and inertial sensors (accelerometers and gyros), but when the former is unavailable (e.g., indoor mapping or outages), and the latter is prone to drift, vision-based navigation offers the natural complement. This is due to the fact that it produces a full six degrees of freedom (6DOF) motion estimate and has lower drift rates than all IMUs with the exception of the most expensive ones (Howard, 2008). Among the available sensors, cameras are affordable and provide rich information on the environment that allows robust and accurate place recognition. The determination of image orientation and localization with respect to a pre-determined 3-D coordinate system is a standard photogrammetric task (Wang et al., 2019), which is often related to structure from motion (SfM) and simultaneous localization and mapping (SLAM) as well as visual odometry (VO). SfM tackles the recovery of both the 3-D scene structure and camera pose from sequentially ordered or unordered image sets. The final structure and camera pose are typically refined with an offline optimization (i.e., bundle adjustment), whose computation time grows with the number of images (Frahm et al., 2010). Conversely, VO focuses on estimating the 3-D motion of the camera sequentially and in real-time. Bundle adjustment can be used to refine the local estimate of the trajectory while SLAM techniques build a map of an unknown environment and localize the platform in the map with a strong focus on real-time operation (Scaramuzza, Fraundorfer). Visual based SLAM (V-SLAM) can be performed by monocular cameras and has been the proposed format by many (Davison et al., 2007; Engel et al., 2014; Mur-Artal et al., 2015). Nonetheless, the scale of the constructed map and estimated trajectory is lost as the distance from the platform to the observed scene cannot be derived from a single view. In contrast with the monocular version, stereo-based V-SLAM utilizes two rigorously connected cameras (Fig. 1) that point to the same direction, allowing to observe depth directly (Mur-Artal, Tardós; Smolyanskiy, Gonzalez-Franco). While stereo vision systems enable to observe depth, its reliability for pose estimation is not necessarily high and depends on multiple properties of both sensor setup and its relation with the observed scene. The ratio between baseline and point distance is commonly used to determine their reliability, but this determination is based on empirical tests rather than actual ones (Paz et al., 2008; Strasdat et al., 2011; Mur-Artal, Tardós). Moreover, vision-based localization depends on the sensor parameters such as focal length and field of view (FOV) angle. While their impact is vastly known, a thorough quantitative analysis on the resulting pose estimation has never been performed. Such an analysis could contribute to designing a system. Airborne platforms, as an example, fly at altitudes of 30-50 meters above the ground where it is reasonable to believe that all extracted features would be at equal range and distant from the platform. Thus, disregarding distant points for pose estimation is not an option.

In this paper, we investigate autonomous pose estimation and evaluate the benefits of stereo-based sensors over monocular ones. For efficient modeling, a novel pose estimation method is introduced. The proposed formulation offers two main advantages over existing ones. First, it allows to use features for pose estimation regardless of the number of cameras they are viewed

*Corresponding author

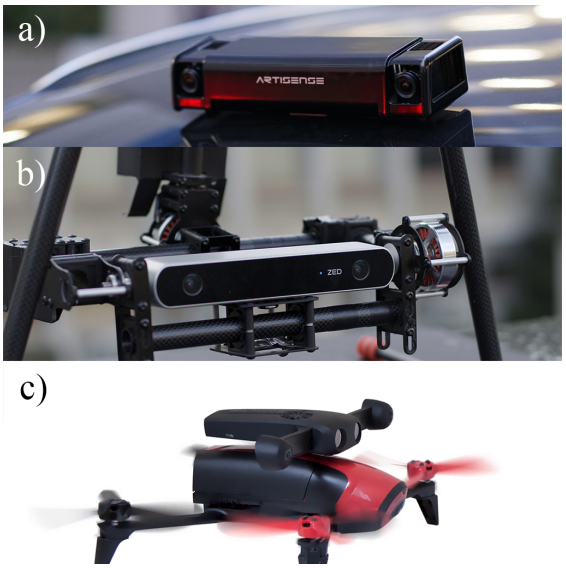

Figure 1: Stereo vision sensors: a) ARTISENSE VINS (Artisense Corporation, 2019), b) ZED stereo camera (StereoLabs, 2019) and c) Parrot S.L.A.M.dunk (Parrot Drones SAS, 2019)

by or their distance from the platform, and it provides a computationally efficient parameter estimation by considering the relative orientation between the sensors as a single entity. This allows reducing the computation demands of bundle adjustment processes and enables efficient integration with Kalman filtering for real-time applications.

2. RELATED WORD

V-SLAM can be performed by a single camera, which is the cheapest and smallest sensor setup. However, as depth is not observable, the scale of both the map and estimated trajectory is unknown. In addition, monocular SLAM suffers from scale drift and may fail if pure rotations are performed during the platform exploration. Using a vision-based stereo camera (i.e., two cameras) resolves all these matters and offers the most reliable V-SLAM solutions Mur-Artal (Tardós).

Existing stereo SLAM systems are mostly keyframe-based (Strasdat et al., 2011) and perform the bundle adjustment computations in local areas (referred as sliding window bundle adjustment) to reduce the scale drift (Scaramuzza, Fraundorfer; Mur-Artal, Tardós). These methods rely on tie-point measurements. Extensive work and many algorithms have been proposed to robustly extract, describe, and match common points, ideally invariant to orientation, scale and illumination changes (Lowe, 2004; Rublee et al., 2011; Leutenegger et al., 2011; Muja, Lowe). Feature extraction and matching are often followed by outlier detection. Once removed, both the keyframe pose and the tie-point 3-D position is estimated and/or updated. This process is repeated for every new pair of keyframes that is introduced until the image acquisition is completed (Fig. 2).

While stereo-vision-based systems allow to observe depth, its reliability for pose estimation is not necessarily high. Civera et al. (2008) was the first to suggest that depth cannot be reliably estimated due to small disparities, and proposed an inverse depth parametrization to distinguish 3-D points which are reliable for localization purposes. Using their formulation, an observed point is excluded from the localization process until it receives a high parametrization value. Paz et al. (2008)

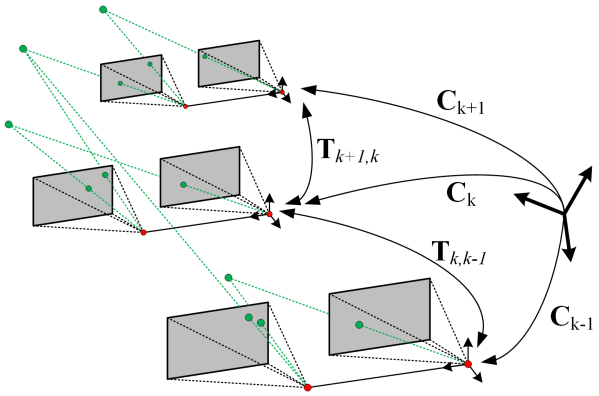

Figure 2: Stereo Visual SLAM process illustration

were the first to propose stereo SLAM method that addressed depth within its localization scheme and showed empirically that points can be reliably triangulated if their depth is less than 40 folds the stereo baseline. This ratio is commonly used in more recent works where it is employed as a threshold value to distinguish between close and far points. Strasdat et al. (2011) used this ratio for full pose estimation, filtering distant points, while Mur-Artal (Tardós) exploited the distant points for rotation computation. Nonetheless, such a distinction cannot be applied in all cases. With airborne platforms, all observed points are equally distant from the sensors, and therefore, all must be included in the localization process. Such localization is common in photogrammetry, and the sole difference is that the length of the baseline and its ratio with respect to the flying altitude is subjected to operational design. In addition, vision-based sensors suffer from a relatively narrow FOV, limiting their ability to observe features for prolonged periods (Herath et al., 2007). As a result, the focal length (and consequently the FOV) determines how fast the platform can turn (Barry et al., 2015). A slight improvement has been made with the introduction of wide-angle lenses. However, these suffer for noticeable lens distortions which also affects the feature extraction and matching process.

3. METHODOLOGY

We consider two rigorously connected cameras with their known pose $\mathbf{c}_i$ and orientation $\mathbf{R}_i$. The overall reference frame is defined with its origin at the middle between the cameras (Fig. 3). The system's orientation is defined as:

$$\hat{\mathbf{x}} = \frac{\mathbf{c}_2 - \mathbf{c}_1}{\|\mathbf{c}_2 - \mathbf{c}_1\|} = \frac{2\mathbf{b}}{\|2\mathbf{b}\|} \tag{1}$$

$$\hat{\mathbf{z}} = \frac{\left(\mathbf{I} - \hat{\mathbf{x}}\hat{\mathbf{x}}^T\right)\left(\hat{\mathbf{z}}_1 + \hat{\mathbf{z}}_2\right)}{\|\left(\mathbf{I} - \hat{\mathbf{x}}\hat{\mathbf{x}}^T\right)\left(\hat{\mathbf{z}}_1 + \hat{\mathbf{z}}_2\right)\|} \tag{2}$$

where $\hat{\mathbf{z}}_{i\in[1,2]} = \mathbf{R}_i \cdot \begin{bmatrix}0 & 0 & 1\end{bmatrix}^T$ represents its optical axis of the i-th camera. With $\hat{\mathbf{y}} = \hat{\mathbf{z}} \times \hat{\mathbf{x}}$, the system orientation becomes:

$$\mathbf{R}_s^r = \begin{bmatrix}\hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}}\end{bmatrix} \tag{3}$$

For airborne platforms the y-axis represents the flight direction while for grounded ones it indicates the 'up' direction. Given

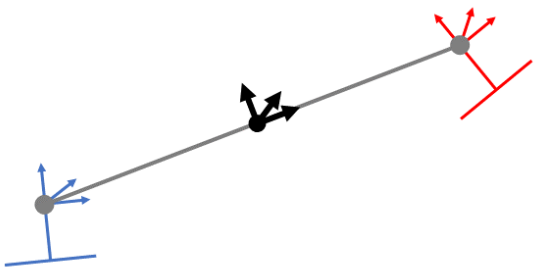

Figure 3: Dual camera system sketch

the reference system, the cameras' relative pose is given by:

$$\begin{aligned} \mathbf{c}_{s_i}^r &= \pm\mathbf{b} \\ \mathbf{R}_{s_i}^r &= \mathbf{R}_s^r \cdot \mathbf{R}_i^T \end{aligned} \tag{4}$$

Similarly, given the exterior orientation of the system, the cameras' pose is defined by:

$$\begin{aligned} \mathbf{c}_{s_i} &= \mathbf{c}_s \pm \mathbf{R}_s\mathbf{b} \\ \mathbf{R}_{s_i} &= \mathbf{R}_{s_i}^r \cdot \mathbf{R}_s \end{aligned} \tag{5}$$

Substituting Eq. (5) into the perspective projection form, the image coordinates, $\mathbf{x}$, of a given ground position, $\mathbf{X}$, in either camera is given by:

$$\begin{aligned} \mathbf{x} &\equiv \mathbf{K}_i\mathbf{R}_{s_i}^T\mathbf{R}_s^T\left(\mathbf{X}-\mathbf{c}_s \mp \mathbf{R}_s\mathbf{b}\right) \\ &\equiv \mathbf{K}_i\left(\mathbf{R}_{s_i}^r\right)^T\left[\mathbf{R}_s^T\left(\mathbf{X}-\mathbf{c}_s\right) \mp \mathbf{b}\right] \end{aligned} \tag{6}$$

where $\mathbf{K}_i = diag\left(-f_i, -f_i, 1\right)$ is the calibration matrix of either camera and f_i is the corresponding focal length. For simplicity, we define: $\mathbf{d} = \mathbf{R}_s^T\left(\mathbf{X}-\mathbf{c}_s\right) \mp \mathbf{b}$. Substituting $\mathbf{d}$ back into Eq. (6) gives:

$$\mathbf{x} \equiv \mathbf{K}\left(\mathbf{R}_{s_i}^r\right)^T\mathbf{d} \tag{7}$$

and the resulting equivalence relation can then be expressed by taking the cross-product:

$$\mathbf{S}_\mathbf{x}\mathbf{K}\left(\mathbf{R}_{s_i}^r\right)^T\mathbf{d} = \mathbf{0} \tag{8}$$

with $\mathbf{S}_\mathbf{x}$ is the skew matrix representation of the vector $\mathbf{x}$. From Eq. (8) we obtain:

$$\begin{aligned} x_i\left(\mathbf{r}_{s_i}^r\right)_{[3]}^T\mathbf{d} &= -f_i\left(\mathbf{r}_{s_i}^r\right)_{[1]}^T\mathbf{d} \\ y_i\left(\mathbf{r}_{s_i}^r\right)_{[2]}^T\mathbf{d} &= -f_i\left(\mathbf{r}_{s_i}^r\right)_{[3]}^T\mathbf{d} \end{aligned} \tag{9}$$

Note that both the relative orientation of the cameras with respect to the defined reference system and their distances from the origin are obtainable via system calibration and Eq. (1) – (4), as well as the cameras' focal lengths. Thus, the remaining unknown parameters in Eq. (9) are the position and orientation of the system (incorporated within $\mathbf{d}$).

3.1 Incorporation into a SLAM scheme

The literature review shows that most SLAM methods employ either local or global bundle-adjustment optimization solution within their workflows. Both types of procedures involve estimating the cameras' pose, rotation, and the 3-D coordinates of the extracted features by minimizing a predefined cost function, which is defined by the reprojection error – the differences between the observed and their back-projected estimated values (Eq. 10).

$$\theta = \underset{\mathbf{c_i},\mathbf{R}_i,\mathbf{X_j}}{\mathbf{argmin}} \sum_{\substack{k\,\in\,\mathcal{K} \\ p\,\in\,\mathcal{P}}} \sum_{q\in\Omega} \left\|\mathbf{x_q} - \mathbf{K_k}\mathbf{R_k}\left(\mathbf{X}_p - \mathbf{c}_k\right)\right\|_\Sigma^2 \tag{10}$$

where $\mathcal{K}$, $\mathcal{P}$, Ω are the sets of keyframes, 3-D points, and their corresponding 2-D image samples involved in the optimization, respectively. For a global bundle adjustment, all keyframes and object points are involved in the computation while for local ones, only a subset of the keyframes and the corresponding visible points are refined.

Optimizing the bundle adjustment problem is achieved by an iterative implementation of the Levenberg-Marquardt algorithm, which solves the minimum of a second-order approximation of Eq. (10) with fixed weights:

$$\delta\hat{\xi}^{(i)} = \left(\mathbf{J}^T\mathbf{W}\mathbf{J} + \lambda \cdot diag\left(\mathbf{J}^T\mathbf{W}\mathbf{J}\right)\right)^{-1}\mathbf{J}^T\mathbf{W}\mathbf{F} \tag{11}$$

where $\delta\hat{\xi}$ is the estimated differential correction of the image orientation parameters and the 3-D coordinates tie-points; $\mathbf{J}$ and $\mathbf{J}^T\mathbf{W}\mathbf{J}$ are Gauss-Newton approximations of the Jacobian and the Hessian of Eq. (10), respectively, $\mathbf{F}$ is the reprojection error vector with respect to the current pose estimates, $\mathbf{W}$ is a diagonal weight matrix of the point samples, and λ is the damping factor. The value of the damping factor varies during the iterations to assure convergence, and is inversely related to the norm of $\delta\hat{\xi}$. The expression $\mathbf{J}^T\mathbf{W}\mathbf{J} + \lambda \cdot diag\left(\mathbf{J}^T\mathbf{W}\mathbf{J}\right)$ is often referred to as the *augmented* Hessian matrix, as a diagonal matrix is being added to the original Hessian.

Decomposing the Hessian matrix into the three sub-matrices: $\mathbf{H}_{11}$, $\mathbf{H}_{22}$ and $\mathbf{H}_{12}$, where the first two are block diagonal and relate to the individual camera parameters and the tie-point coordinates, respectively; and the third ($\mathbf{H}_{12}$) maps the relations between points and cameras, and using the Schur complement we can write:

$$\left(\mathbf{H}_{11} - \mathbf{H}_{12}\mathbf{H}_{22}^{-1}\mathbf{H}_{12}^T\right)\delta\xi_{cam} = \mathbf{u}_1 - \mathbf{H}_{12}\mathbf{H}_{22}^{-1}\mathbf{u}_2 \tag{12}$$

$$\delta\xi_{pts} = \mathbf{H}_{22}^{-1}\left(\mathbf{u}_2 - \mathbf{H}_{12}^T\delta\xi_{cam}\right) \tag{13}$$

where $\mathbf{u}_1$ and $\mathbf{u}_2$ are the respective error vectors for the camera and tie point related parameters, $\delta\xi_{cam}$ and $\delta\xi_{pts}$, respectively (Lourakis, Argyros). We apply the Cholesky factorization for solving the cameras pose parameters. Notably, with the proposed representation, instead of solving the two cameras that form the stereo-setting, only a single pose for the two is estimated. Thus, the number of unknowns is reduced by half. This, in turn improves the computational efficiency by a factor of four.

Keyframe Insertion – Considering the amount of imagery data acquired, using all information in the localization process is impractical, as the cameras operate in faster rates than the platform's movement. Therefore, only a reduced set of keyframes is relevant for evaluating the platform's pose and orientation. New keyframes are introduced when sufficient movement is

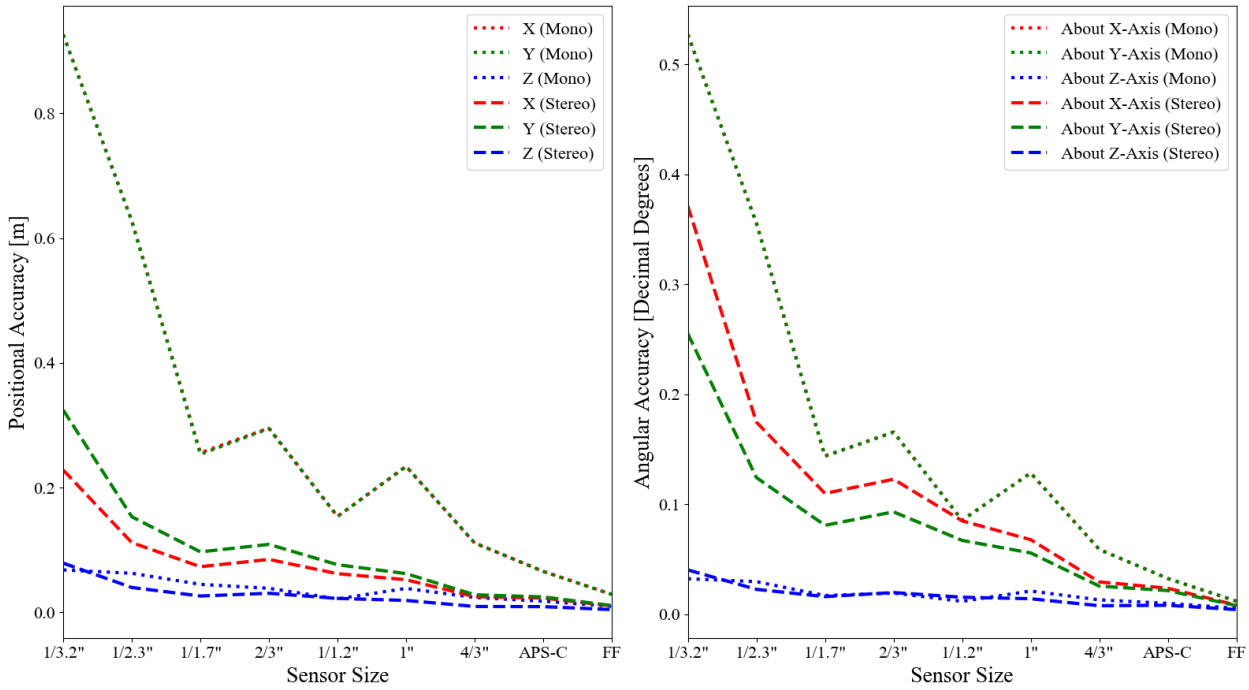

Figure 4: Sensor size implications on pose accuracy for monocular and stereo vision sensors

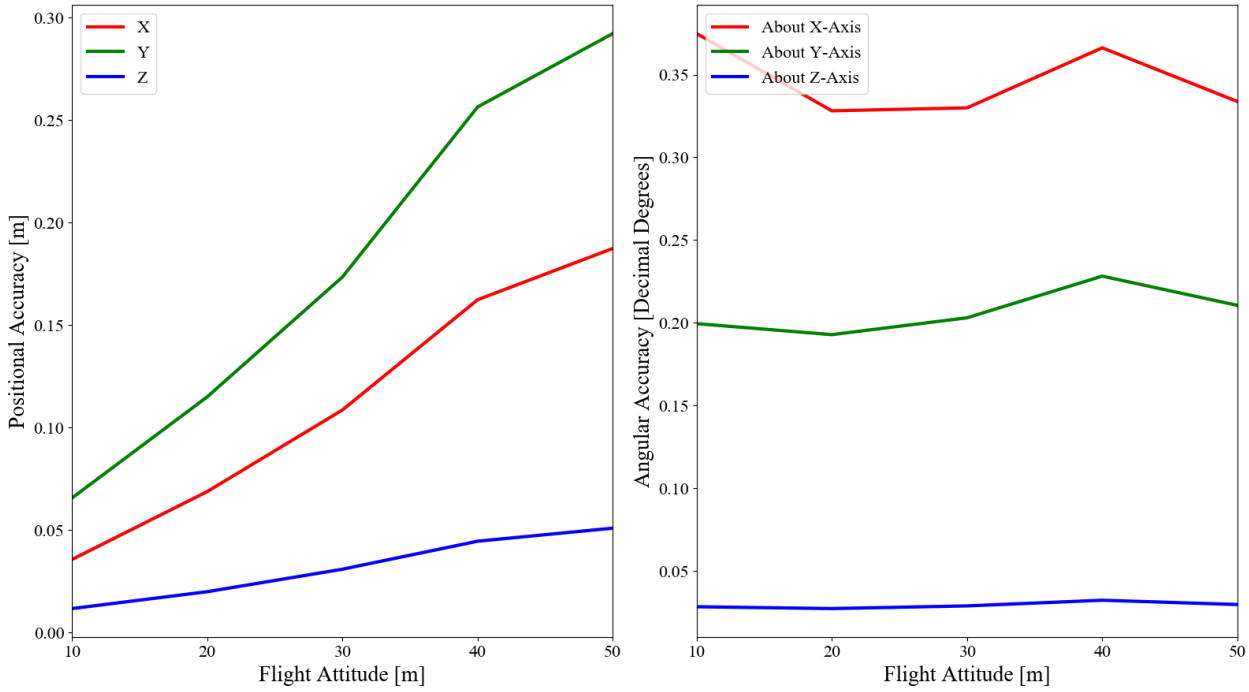

Figure 5: Altitude implications on pose accuracy for stereo vision sensors

recorded. Current methods translate this to maintain a sufficient number of reference points whom distance from the cameras is small (Strasdat et al., 2011; Mur-Artal, Tardós). In the present case the criterion would be maintaining sufficient overlap between the keyframes so that the predefined pose accuracy would be reached. This translates to the introduction of new frames at constant time intervals, which are determined by the flying altitude and velocity of the platform.

Keypoint Matching – As the relative orientation between the cameras is known, the keypoint matching is partitioned into finding matches between images taken at different instances and finding matches between the individual stereo pairs. For the matching between different instances, the fast library for approximate nearest neighbors (FLANN; Muja, Lowe) is used for querying. For the matching within the individual stereo-pairs, an efficient search along the epipolar lines would suffice.

4. ANALYSIS

In order to test the merit in using a stereo-sensor vs. a monocular one, we evaluate the contribution along four avenues, including: the impact of the sensor size, the baseline, the operational

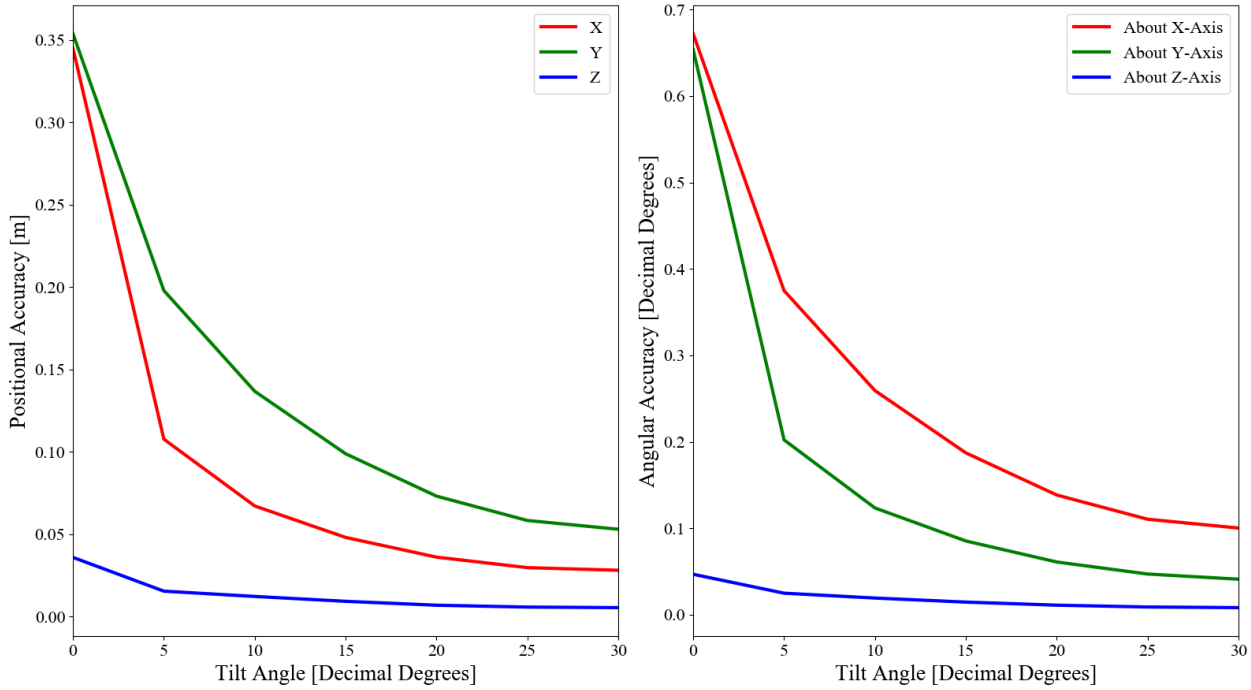

Figure 6: Effect of angle between the sensors' optical axes on pose accuracy

Camera Type	Focal Length [mm]	Sensor Size	FOV angle [°]
First Person View (FPV)	2.0 – 2.8	1/3" – 1/1.8"	100 – 190
Extreme	3 – 50	1/3" – 4/3"	55 – 150
Compact	20 – 80	1/3" – 4/3"	55 – 120
DSLR	$\geq$ 8.5	APS-C	$<$ 170
Full Frame	$\geq$ 8.5	35mm	$<$ 170

Table 1: Existing vision sensors

altitude, and the tilt angle, on the accuracy. As the sensor size is one of the first design parameters it is evaluated first. For cameras, there is an ever increasing variety (Table 1), where most platform designs tend to smaller cameras as they are affordable and weight less. Such consideration does not necessarily takes into account the pose estimation accuracy. To evaluate the sensor size implications on the derived pose estimates we consider an operational altitude of 50 m above ground and compare a stereo-setting to a monocular case. Evaluation shows that the pose estimation (location and orientation) improved as a function of the sensor size in use for both stereo and monocular scenarios, but the stereo one outperforms the monocular solution (Fig. 4). For a small sensor size, the contribution of the stereo setting was up to nearly five-fold in positional accuracy. The improvement in accuracy as the dimensions of the sensor increase are more moderate when using the stereo-setting than that of the monocular case. Only when using a full frame sensor the accuracy of the monocular solution and the stereo one are the same. Clearly, the cost of a full frame camera is much higher than that of a first-person-view (FPV) solution. In sum, our results suggest that stereo vision sensors allow reducing the camera size with minimal impact on the quality of the derived position and orientation of the platform. Assuming a fixed focal length, an increase in size of the sensor relates also to an increase in field of view, which in turn allows triangulating the platform's location by a using more reference data.

Further examination evaluated the performance of the stereo-setup over different operational altitudes (Fig. 5). The results show that while location accuracies decreased with the increase in altitude, no significant change in the quality of the rotational parameters estimates was observed. These results are in agreement with others who have demonstrated that points in greater distances are useful only for orientation estimation.

Evaluation of the baseline impact on the accuracies, shows little contribution if any. This is an expected outcome as the base-to-height ratio of such platform is negligible to be meaningfully affected by a change in the baseline. Our earlier experiments were using a 10 cm baseline, about the dimension that one would expect with such systems, yet outperforming the monocular solution.

While the baseline between the cameras received much attention in the literature (e.g., Engel et al., 2015; Mur-Artal, Tardós), the tilt angle between them (the angle between the optical axes), receive only little. This has mostly to do with the use of the stereo-setting for depth extraction, but performance of the pose estimation using only a single view. As the model allows examining the direct impact of all system parameters, we test what the contribution of a tilted setting is. Clearly, the increase of the tilt angle between the two cameras increases the field of view (Fig. 7). The results (Fig. 6) show that the effect of the field of view on the positional accuracy is dramatic, and even with a modest 10° inclination the positional accuracy is $\sim$ 10 cm and 2 cm in altitude. The difference in accuracy between the $x-$ and y-directions has to do with the stereo-system alignment, which is orthogonal to the platform motion direction along the y-axis. A 15° angle yields a sub-decimeter accuracy in all axes and an angular accuracy of 12'. Thus, we conclude that even a relatively modest change of the tilt angle is sufficient to secure accurate pose parameters estimates. Further increase of the angle has a relatively moderate contribution, and we also note that the magnitude of the tilt angle should also consider the decrease in the overlap between the images taken at the same instance and the obliqueness of the images. Obliqueness may affect the quality of the extracted keypoints and reduced the overlap between the two frames of the stereo sensor into two

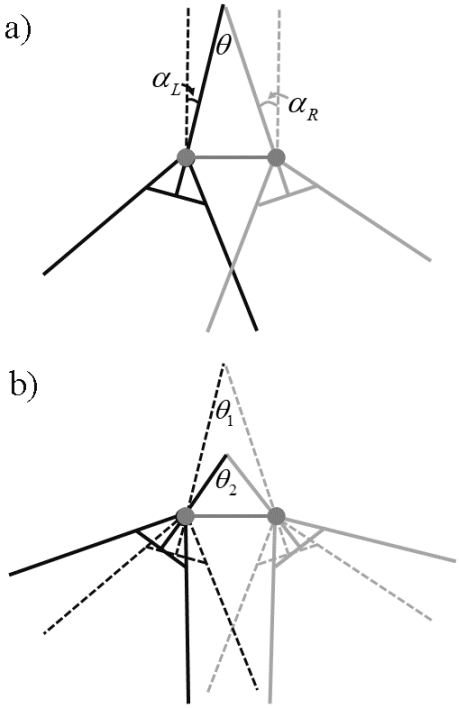

Figure 7: Field of view angle of a stereo vision sensor in relation to the tilt angles (a) and broadening by them (b).

monocular imaging setting. This will limit the mutual ground coverage of both.

5. CONCLUSIONS

This paper we presented an alternative model for efficient pose estimation of a mobile platform using a stereo vision sensor. The proposed formulation not only reduces the computational complexity of visual-based localization, but also provides a framework for system performance investigation. Using the proposed formulation and by taking advantage of the extended field of view from both cameras, a five-fold improvement in localization is reached for standard operational altitudes of 50 m, from ∼0.4-0.5 m with a single sensor to less than 0.1 m. From a system configuration aspect, Our analysis shows that even a slight modification in cameras' directions improves the positional accuracy. Furthermore, such improvement is reached using cameras with reduced sensor size, which are more affordable. Hence, it is shown that not only does a dual-camera setup improves pose estimation, but it also enables to use smaller sensors and reduce the overall system cost.

REFERENCES

Artisense Corporation, 2019. Visual-Inertial Navigation System (VINS) DevKit. `https://www.artisense.ai/localization`. Last Accessed: October 2019.

Barry, A. J., Oleynikova, H., Honegger, D., Pollefeys, M., Tedrake, R., 2015. Fast onboard stereo vision for uavs. *Vision-based Control and Navigation of Small Lightweight UAV Workshop, International Conference On Intelligent Robots and Systems (IROS)*.

Civera, J., Davison, A. J., Montiel, J. M., 2008. Inverse depth parametrization for monocular SLAM. *IEEE Transactions on Robotics*, 24(5), 932–945.

Davison, A. J., Reid, I. D., Molton, N. D., Stasse, O., 2007. MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 29(6), 1052–1067.

Engel, J., Schöps, T., Cremers, D., 2014. LSD-SLAM: Large-scale direct monocular SLAM. *European Conference on Computer Vision (ECCV)*, Springer, 834–849.

Engel, J., Stückler, J., Cremers, D., 2015. Large-scale direct SLAM with stereo cameras. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 1935–1942.

Frahm, J.-M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Dunn, E., Clipp, B., Lazebnik, S. et al., 2010. Building Rome on a cloudless day. *European Conference on Computer Vision (ECCV)*, Springer, 368–381.

Gerke, M., Nex, F., Jende, P., 2016. Co-Registration of Terrestrial and UAV-Based Images – Experimental Results. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-3/W4, 11–18.

Herath, H., Kodagoda, S., Dissanayake, G., 2007. Stereo vision based SLAM: Issues and solutions. *Vision Systems: Applications*, ITECH.

Howard, A., 2008. Real-time stereo visual odometry for autonomous ground vehicles. *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 3946–3952.

Leutenegger, S., Chli, M., Siegwart, R., 2011. BRISK: Binary robust invariant scalable keypoints. *2011 IEEE international conference on computer vision (ICCV)*, IEEE, 2548–2555.

Lourakis, M. I., Argyros, A. A., 2009. SBA: A software package for generic sparse bundle adjustment. *ACM Transactions on Mathematical Software (TOMS)*, 36(1), 2.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 91–110.

Muja, M., Lowe, D. G., 2009. Fast approximate nearest neighbors with automatic algorithm configuration. *International Conference on Computer Vision Theory and Applications (VISAPP '09)*, 2, 331–340.

Muja, M., Lowe, D. G., 2012. Fast matching of binary features. *2012 Ninth Conference on Computer and Robot Vision*, IEEE, 404–410.

Mur-Artal, R., Montiel, J. M. M., Tardos, J. D., 2015. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5), 1147–1163.

Mur-Artal, R., Tardós, J. D., 2017. ORB-SLAM2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5), 1255–1262.

Parrot Drones SAS, 2019. Parrot S.L.A.M.dunk. `https://developer.parrot.com/docs/slamdunk/`. Last Accessed: October 2019.

Paz, L. M., Piniés, P., Tardós, J. D., Neira, J., 2008. Large-scale 6-DOF SLAM with stereo-in-hand. *IEEE Transactions on Robotics*, 24(5), 946–957.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G. R., 2011. ORB: An efficient alternative to SIFT or SURF. *2011 International Conference on Computer Vision (ICCV)*, IEEE, 2564–2571.

Scaramuzza, D., Fraundorfer, F., 2011. Visual Odometry [Tutorial]. *IEEE Robotics Automation Magazine*, 18(4), 80-92.

Smolyanskiy, N., Gonzalez-Franco, M., 2017. Stereoscopic First Person View system for Drone navigation. *Frontiers in Robotics and AI*, 4, 11.

StereoLabs, 2019. ZED stereo camera. `https://www.stereolabs.com/zed/`. Last Accessed: October 2019.

Strasdat, H., Davison, A. J., Montiel, J. M. M., Konolige, K., 2011. Double window optimisation for constant time visual SLAM. *2011 International Conference on Computer Vision (ICCV)*, 2352–2359.

Wang, X., Rottensteiner, F., Heipke, C., 2019. Structure from motion for ordered and unordered image sets based on random k-d forests and global pose estimation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 147, 19–41.