Scaled Monocular Visual SLAM

Georges Younes University of waterloo, ON, Canada
Daniel Asmar American University of Beirut, Beirut, Lebanon
John Zelek University of waterloo, ON, Canada
David Abou Chacra University of waterloo, ON, Canada
Henry Leopold University of waterloo, ON, Canada
Jeremy Pinto University of waterloo, ON, Canada
Nolan Lunscher University of waterloo, ON, Canada

Abstract

The fundamental shortcoming underlying monocular-based localization and mapping solutions (SfM, Visual SLAM) is the fact that the obtained maps and motion are solved up to an unknown scale. Yet, the literature provides interesting solutions to scale estimation using cues from focus or defocus of a camera. In this paper, we take advantage of the scale offered by image focus to properly initialize Visual SLAM with a correct metric scale. We provide experiments showing the success of the proposed method and discuss its limitations.

1 Introduction

Simultaneous Localization and Mapping (SLAM) is an attractive alternative to motion estimation when GPS cannot be relied on. In the absence of an infrastructure, SLAM solutions are considered indispensable for robot navigation. While SLAM solutions have been proposed for different types of sensors, the employment of a camera as the sole sensory input—systems known as visual SLAM— established its sovereignty in applications where size, weight and power consumption are deciding factors. Visual SLAM branches horizontally into two main categories: stereo systems and monocular systems. While the former offer plenty of advantages, they fall short when compared to the cost and flexibility of monocular methods[1]. In the wake of the increasing interest of the research community in monocular Visual SLAM, many solutions were recently devised[2] based on the skeleton put forward in the pioneering work of Klein and Murray, dubbed PTAM (Parallel tracking and Mapping) [3]. Simply put, PTAM can be stripped down to five core modules, of particular importance to this work, is the initialization module, which was approached differently by different groups. For example, in Davison and Reid's MonoSLAM[4], system initialization required the camera to be placed at a known distance from a planar scene, and SLAM was initialized with the distance keyed in by the operator. PTAM, suggested the usage of the five-point algorithm [5] to estimate and decompose a Fundamental matrix into an assumed to be non-planar initial scene. PTAM initialization was later changed to the usage of a Homography [6],where the scene is assumed to be composed of 2D planes. Other systems include the work of Forester et al. [7] who adopted a Homography, or that of Tan et al.[8] and Herrera et al. [9] who used an Essential matrix [10], where the scene is assumed non-planar. With the exception of MonoSLAM, all the suggested methods suffered from degeneracies when their implied assumption of the scene was violated. To address the issue, Mur-Artal et al. [11] employed both methods in parallel and suggested a metric to elect either one when a degenerate case is detected in the other. SchÃűps and Cremers [12] suggested a randomly initialized scene's depth from the first viewpoint that is later refined through measurements across subsequent frames. Limited by the capabilities of single cameras to generate bearing-only measurements, a fundamental limitation of all monocular Visual SLAM solutions is that the actual scale of the scene cannot be recovered accurately during initialization. This means that, without additional information, the relationship between image coordinates and corresponding 3D point coordinates can only be determined up to an unknown scale λ.

The work in this paper aims to diminish the above limitation by suggesting a novel initialization technique for Visual SLAM that, unlike MonoSLAM requires no human input. The gist of the solution is to determine scale of a scene using depth from focus. More specifically, during initialization, the camera is moved normally to the scene in search of the image that is most focused. This is possible by performing an offline pre-calibration of the camera, where for a given camera focal distance, we determine the corresponding scene depth producing maximum image focus. Although the system is very sensitive to motion rotation, experiments demonstrate the success of the proposed technique. The remainder of this paper is structured as follows. Section 2 surveys depth from focus methods. Section 3 describes our adaptation and implementation of depth from focus to initialize and estimate the scale of a visual SLAM system. Section 4 presents the experiments and section 5 discusses the obtained results. The paper concludes in Section 6.

2 Depth from focus

While traditional visual SLAM concerns itself with depth estimation from parallax, research in optics suggest two methods capable of recovering depth from images that do not exhibit parallax; namely, depth from focus and depth from defocus. Depth-from-Defocus (DfD) provides a solution to estimate the depth of a scene by measuring the blurriness of objects. On the other hand, Depth-from-Focus (DfF) recover depth by searching for the state of the imaging system for which the object is in-focus in the image plane. This can be achieved by either (1) varying the distance between the lens and the imaging sensor or by (2) varying the distance from the lens to the observed scene. Objects that are at the focal plane will result in maximum focus on the camera sensor. Any change from the focal plane results in a blurred representation of the scene. In the first technique, for each scene the focus of camera is varied until the image is in focus. Pucihar and Coulton [13], provide such a solution, which requires an offline calibration resulting in a lookup table relating focus-to-depth. Allowing the camera to change its focus during a visual SLAM session is ruinous because the intrinsic camera calibration parameters, vital to achieve acceptable tracking performance, varies with the focus and would lead to tracking failure. Furthermore, not all cameras retain an auto-focusing mechanism nor allow access to their drives. Suwajanakorn et al. [14] recently suggested an algorithm capable of recovering depth from focus using an uncalibrated camera; however, their suggested pipeline requires twenty minutes of processing, which is intractable for real-time operations.

In this work, we adopt the second approach by moving the camera during initialization in the direction normal to the scene along its optical axis and estimate depth corresponding to the most in focus image. For this method to succeed, an appropriate focus measure operator is a critical in ensuring accurate depth estimation. A wide variety of algorithms [15] have been used to measure the degree of focus of image patches or the image as a whole. Given their real-time requirements, Visual SLAM implementations pose constraints on the amount of allowable processing time for each frame, Therefore, a simple, fast, and relatively accurate focus operator is used, consisting of first extracting the Laplacian eq.1

$$\nabla^2 I(x,y) = \frac{\partial^2 I(x,y)}{\partial x^2} + \frac{\partial^2 I(x,y)}{\partial y^2} \tag{1}$$

and then summing the Laplacian over a window as in eq.2; a step necessary to help deal with poorly-textured surfaces.

$$FM(x_0,y_0) = \sum_{(x,y)\in\sigma(x_0,y_0)} \nabla^2 I(x,y) \tag{2}$$

where $\sigma(x_0,y_0)$ is the support window chosen as the 24x24 pixel patch centered at $I(x_0,y_0)$. The total focus measure of the image is then found using eq.3

$$F = \frac{1}{n} \sum_{(x,y)\in I} FM(x,y)^2 \tag{3}$$

where n is the number of pixels in the image. Finally, a moving average filter is employed to attenuate the effect of noisy measurements in the focus measure operator.

3 Depth from focus in Visual SLAM

A camera moving longitudinally along a path toward a planar scene exhibits maximum focus when the distance between the camera and the scene is equal to its focal plane. Figure 1 illustrates how the focus measure of an image is expected to vary with the distance from the scene.

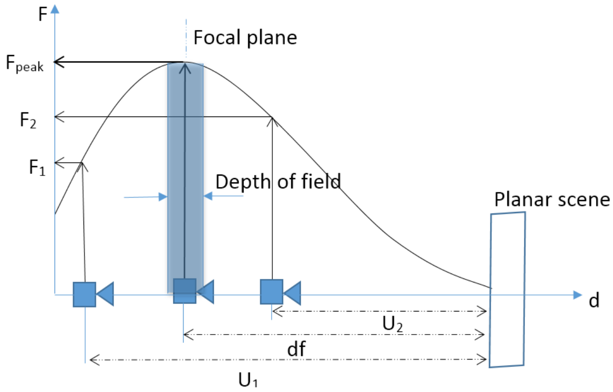

Fig. 1: Focus measure profile vs change in distance between camera and scene.

Therefore, what is sought in this work is an algorithm that is capable of determining, among a range of images, the image exhibiting maximum focus F_{peak} for a given scene. To make this possible, before initializing Visual SLAM, the system would have to be calibrated in order to determine, for a given focal length, the depth corresponding to maximum focus. The proposed calibration is performed offline, by fixing either the camera or the planar scene and moving the other longitudinally away, while monitoring the focus measure response. When a peak in the focus measure response is recorded, the corresponding distance from the camera is registered as the focal plane's distance.

Once the focal plane's distance is known through the calibration process, it can be used to initialize any Visual SLAM system, assuming the observed scene during initialization is planar as is the case in most Visual SLAM implementations. In this paper, we test the technique on PTAM as a replacement for its default initialization using Homographies. Once the system is started, tracking and mapping are set to an idle state, while the user is asked to move forward and backward towards a planar scene. The focus measure is then recorded automatically at every frame and the one corresponding to the peak focus measurement is registered as the first keyframe in the map.

The pose of the keyframe is represented as a rigid body transformation $\in$ SE (3). Initially, the pose $E_{1,w}$, is assigned to the 4×4 identity matrix. FAST features [16] from the 0^{th} pyramid level are then extracted and their 3D coordinates are initialized as in eq.4

$$[X',Y',Z',1]^T = E_{1,w}[\frac{P_x}{D},\frac{P_y}{D},\frac{1}{D},1]^T \qquad (4)$$

where $D = \frac{1}{focalplanedistance}$, hardcoded into the system beforehand through the camera calibration process. P_x and P_y are pixel coordinates of the extracted features projected onto a normalized image plane using the radial distortion model of [17]. Similar to PTAM, the mean of the 3D features is then elected to serve as the world coordinate frame and the entire map is then transformed accordingly. Once the initialization procedure is complete, the visual SLAM system resumes its regular tasks, performing camera tracking and scene mapping.

The initialization here is considerably different from the traditional methods of Visual SLAM, in which the user is required to trigger the system, move the camera a distance that is ad-hoc, and then trigger the system again once sufficient parallax is achieved. In what we are proposing, no human intervention is required; the camera is moved tangent to its optical axis until the system automatically initializes. This type of motion is more natural than a lateral one, especially for mobile platforms such as Unmanned Aerial Vehicles (UAV) with a downward looking camera, or for nonholonomic land vehicles equipped with forward looking cameras.

4 Experiments and Results

Our proposed method was implemented in PTAM [3] and tested on a laptop with an Intel Core i7-4710HQ 2.5GHZ CPU, 16 GB memory; no GPU acceleration was used. As table 1 shows, the computational cost for each focus measurement required at every frame 5.4 ms; once the focused frame is found, our proposed initialization requires 8.5 ms to kick-start the system, in contrast to the default Homography initialization of PTAM that requires on average 250 ms. Furthermore, our proposed initialization does not yield multiple solutions in contrast to the Homography estimation that in some cases may be degenerate or return ambiguous results.

Table 1: Computational cost for initialization

Operation	Time(ms)
Focus measurement	5.4
Proposed initialization module	8.5
Homography initialization	250

To test our system, two experiments were conducted. During our experiments, motion was only in the vertical direction, therefore ground truth was collected manually, by measuring the actual distance between the lens and the scene.

4.1 Experiment 1

Experiment 1 consisted of fixing the camera on a rig that can move in a single direction normal to a planar scene. This was necessary to validate the theory and test its application to PTAM in a controlled environment The camera was focused and calibrated beforehand at a distance of 23 cm. Within the controlled confines of a fixed rig, the experiment was repeated 31 times, either moving towards the planar scene from a starting position of 30 cm or moving away from the scene with a starting position of 10 cm, at a constant rate. The actual distance between the camera and the scene, at which our proposed method initialized the system, was then recorded.

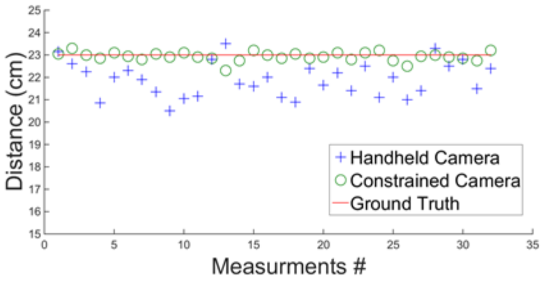

Fig. 2: Experiment 1- System initialization.

The obtained results are shown on figure 2 as circles; with a mean of 22.93 cm and a standard deviation of 0.2 cm, they demonstrate the accuracy and precision of our initialization module. Next, the same experiment was repeated but this time the hand-held camera was free to move in 6D. Nevertheless, the user was asked to avoid high acceleration movements and tilting as much as possible. In this experiment, the objective was to study the impact of factors such as camera orientation changes and motion blur, induced by human interference, on the initialization quality of PTAM.

4.2 Experiment 2

The second experiment consists of initializing PTAM using our method and then recording its camera pose measurements and compare them to the ground truth for both the fixed rig and handheld cases. While this experiment can be easily performed in 6D, for visualization purposes, it was conducted along a single dimension. After initialization, the scene was explored, by inserting keyframes with enough parallax between them, to ensure a good baseline for feature triangulation to take place, before returning to the experiment's configuration of motion along a single direction. The recorded paths are shown in figures 3a and 3b. They show that our initialization module was able to snap to the actual scene's scale by

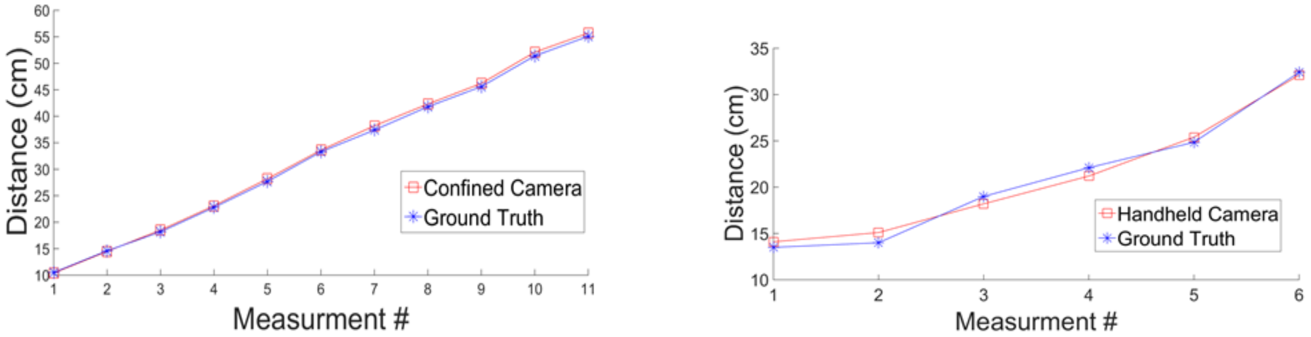

Fig. 3: Experiment 2- Path recorded by PTAM initialized with our method.

yielding pose estimates with an RMSE of 0.49 cm in the confined camera case and 0.62 cm in the handheld version of the experiment.

5 Discussion

The accuracy and repeatability reported in Experiment 1 proves the viability of the suggested initialization under the constraints, where the camera moves normally to the planar scene; however, as the constraints are relaxed through the hand-held version of Experiment 1, where the camera's motion is subject to human interference, the results varied. The decrease in accuracy and repeatability, in the hand-held version, can be traced back to several factors namely, (1) motion blur caused by jittering and fast motions of the camera, (2) rotation deviations from the normal of the scene due to camera handling errors. To reduce the effects of jittering and motion blur on the system, a down sampled representation of the image may be used to estimate the focus measure, at the expense of decreasing its sensitivity. Whereas the second source of error is tightly linked with the camera's motion assumption of moving along its optical axis normal to the observed scene. As the camera is tilted at an angle with the normal to the plane, the reported distance at which the peak focus measure would fall shorter than the actual focal plane distance, which explains why the mean of the hand held version of Experiment 1 was shifted below the actual value of the focal plane.

6 Conclusion

In the course of this work we have presented an adaptation of depth from focus methods to suggest a novel visual SLAM initialization procedure, capable of initializing the system and recovering its accurate scale using a single frame, captured by a monocular camera moving longitudinally towards/away from a planar scene. We proved the accuracy and precision of our initialization procedure through a controlled experiment and highlighted, in another experiment, the challenges that faces its deployment for handheld cameras, mainly motion blur and rotations.

References

[1] Lim, H., Lim, J., Kim, H.J., 2014. Real-time 6-DOF monocular visual SLAM in a large-scale environment, in: Robotics and Automation (ICRA), IEEE International Conference on, pp. 1532–1539.

[2] G. Younes, D. Asmar, and E. Shammas, "A survey on non-filter-based monocular Visual SLAM systems," *ArXiv e-prints* (2016).

[3] Klein, G., Murray, D., 2007. Parallel Tracking and Mapping for Small AR Workspaces. 6th IEEE and ACM International Symposium on Mixed and Augmented Reality , 1–10.

[4] Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O., 2007. MonoSLAM: real-time single camera SLAM. Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions on 29, 1052–67.

[5] Nistér, D., 2004. An efficient solution to the five-point relative pose problem. Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions on 26, 756–77.

[6] Faugeras, O., Lustman, F., 1988. Motion and structure from motion in a piecewise planar environment. International Journal of Pattern Recognition and Artificial Intelligence 02, 485–508.

[7] Forster, C., Pizzoli, M., Scaramuzza, D., 2014. SVO : Fast Semi-Direct Monocular Visual Odometry, in: Robotics and Automation (ICRA), IEEE International Conference on.

[8] Tan, W., Liu, H., Dong, Z., Zhang, G., Bao, H., 2013. Robust monocular SLAM in dynamic environments. 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) , 209–218.

[9] Herrera, D., Kannala, J., Pulli, K., Heikkila, J., 2014. DT-SLAM: Deferred Triangulation for Robust SLAM, in: 3D Vision, 2nd International Conference on, IEEE. pp. 609–616.

[10] R. Hartley. In Defence of the 8-point Algorithm. In 5^{th} *IEEE Trans. Pattern Anal. Mach. Intell., 19(6), 580-593. http://doi.org/10.1109/34.601246*

[11] Mur-Artal, R., Montiel, J.M.M., Tardos, J.D., 2015. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. IEEE Transactions on Robotics PP, 1–17.

[12] Engel, J., Schöps, T., Cremers, D., 2014. LSD-SLAM: Large-Scale Direct Monocular SLAM, in: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014 SE - 54. Springer International Publishing. volume 8690 of *Lecture Notes in Computer Science*, pp. 834–849.

[13] K. Čopič Pucihar, P. Coulton, 2011. Estimating scale using depth from focus for mobile augmented reality, Proceedings of the 3rd ACM SIGCHI Symposium on Engineering Interactive Computing Systems, pp. 253–258.

[14] S. Suwajanakorn, C. Hernández,S. Seitz, 2015, Depth from focus with your mobile phone, CVPR.

[15] S. Pertuz, D. Puig, M.A. Garcia, 2013, Analysis of Focus Measure Operators for Shape-from-focus,Pattern Recogn., 46(5), Elsevier Science Inc., pp. 1415–1432.

[16] Rosten, E., Drummond, T., 2006. Machine Learning for High-speed Corner Detection, in: 9th European Conference on Computer Vision - Volume Part I, Proceedings of the, Springer-Verlag, Berlin, Heidelberg. pp. 430–443.

[17] F. Devernay O. Faugeras,2001, Straight Lines Have to Be Straight: Automatic Calibration and Removal of Distortion from Scenes of Structured Enviroments, Mach. Vision Appl., Springer-Verlag New York, Inc., 13 (1), pp. 14–24.