Position estimation of small UAV from monocular camera by using matched features

Hariom Dhungana, Francesco Bellotti, Riccardo Berta, Alessandro De Gloria
DITEN, University of Genova, 16145, Genova, Italy
{hariom.dhungana@elios.unige.it, franz@elios.unige.it, berta@elios.unige.it, adg@elios.unige.it}

Abstract. This paper focuses on position estimation of a small unmanned aerial vehicle (UAV) using a monocular camera. Features from Accelerated Segment Test (FAST) descriptors are used as a matched pattern to estimate differential change in position of the UAV. Visual Simultaneous Localization and Mapping (V- SLAM) is a probabilistic filter based method and a prominent real time positioning method in robotics. V-SLAM performs drift free tracking of the pose of a UAV on long run but the prediction states has limited certainty because of using sparse number of features used in real time position estimation. Visual Odometry (VO) is a deterministic positioning method and is more accurate to estimate the relative UAV position from adjacent frames without a persistent map. VO gives high drift error on the long run because of the accumulation of drift error at each frame transformation. Bundle Adjustment (BA) and loop closure are two error optimization techniques to reduce drift error in VO. Due to limited computation resources available in the small scale UAV the optimization techniques are not appropriate in the small UAV positioning. In this work, VO with fractional V-SLAM is proposed to reduce the drift error on position estimation from matched features. The obtained results show the positioning estimation from proposed method works in an outdoor environment and that over performs than VO and V-SLAM.

Keywords: Positioning, UAV, visual odometry, visual SLAM

1 Introduction

Small UAVs are used in various application scenarios such as search and rescue, monitoring and security, tracking and observation. Due to the applicability of such UAVs in civil applications [1], their importance grows day by day. Traffic monitoring, fire detection, pipeline/power line inspection, industrial inspections are the common civil applications that are performed by UAVs. The performance of application scenarios depends on the accuracy of the UAV position. Coordinated search, virtual views and image stitching are usual activities of the small UAV in an outdoor environment. In order to cover the interested area in a short duration by multiple UAVs, there is a need for coordination. The coordinated search needs sharing of position information to neighboring UAVs [2]. Quick generated virtual views and 3D reconstruction helps in the analysis of damages from an earthquake. Minimum number of images captured by the UAVs to cover interest region makes the swiftness in 3D reconstruction [3]. Appropriate position selection of the camera gives the full coverage with minimum number of images, which helps quick formation of 3D map. Mosaicking of images gives a big picture of visual information in a single frame. Image mosaicking suffers from metadata error and camera orientation error [4]. Accurate location and orientation information of camera helps to reduce metadata error in image stitching. These are the motivation to get accurate positioning of the UAVs.

Defining an UAVs location and orientation with some reference in 3D space is called positioning and requires six parametric values to represent the complete information. Positioning can be defined in two ways: relative and absolute positioning. The relative positioning deals with comparing to nearby objects or periphery as reference and is commonly used in human perception. The absolute positioning is defined with respect to some global reference. Remarkable achievement of the computational power in small circuits makes it easy to implement visual based positioning in small UAVs. Since, there is already a camera in a small UAV; it can be used as a positioning sensor. A camera is a passive device in terms of radio interference so useful in passive observation. Images and videos are the output data from the camera. Visual positioning is the process of calculating differential change in position of the UAV by analyzing repeated patterns that appeared on the consecutive image frames. Our assumptions for visual position estimation are availability of sufficient illumination in the environment; sufficient scene overlap between successive frames and dominance of static scene over moving objects. Position sensing is a fairly mature field; however, there is a need for research for cost minimization, infrastructure reduction, scalability improvement and flexibility on implementation as stated in [5].

Davisons monocular SLAM [6] and Nisters VO [7] is perhaps the parent of our work, and were far ahead of their time. Our contribution in this paper is to adapt the related work of Extended Kalman Filter (EKF) based monocular SLAM [6], VO [7], inverse depth representation of features [8] and camera centric coordinate system [9] to estimate the position of the small UAV. Furthermore, we have incorporated the combination of Random sample consensus (RANSAC) with EKF as in [10] to speed up the capability to detect and reject a high number of spurious features.

2 Related works

The real time VO with low delay was proposed in [7] using the Harris corners. The disparity between the matched features depends on the speed of the camera and is tested in different ranges. Minimum five matched points in consecutive image frames are needed to estimate the six parameters of position from a calibrated camera. In practice, more than five points are used for the motion estimation to get robust and accurate. More than eight points matching leads to an overdetermined system to solve in the least squares sense   and provides a degree of robustness to noise [11]. In [12], Scaramuzza and Siegwart demonstrated real-time algorithms for calculation ego-motion of a ground vehicle relative to the road using an omnidirectional camera. VO on the ground plane using only one feature correspondence was tested in [13]. Computation time per frame was less than one millisecond and the number of iterations for outlier removal was seven. Circular motion model of two Degrees of Freedom (DoF) was used for camera motion. Although the method was fast and efficient, it was insufficient to represent all position parameters of the UAV from single feature correspondence. VO system based on two-frame estimates of instantaneous relative motion can work in constant time, but will inevitably exhibit drift because of accumulation of small errors in the inter frame motion estimates. Bundle adjustment [14] and loop closure are two error optimization techniques to reduce drift error in VO. BA reduces the drift compared to two view VO because it incorporates constraints between several frames. The computational complexity of BA increases cubically $\boldsymbol{O((qN + lm)^3)}$ with number of parameter involve in BA, where $\boldsymbol{q}$ represents number of parameters to describe feature points, $\boldsymbol{N}$ represents number of feature points, $\boldsymbol{l}$ represents number of camera poses and $\boldsymbol{m}$ represents number of parameters to describe camera poses. For this reason, BA   is not feasible in the small UAV positioning. The loop closer is not always available in outdoor UAV missions.

EKF was first successfully applied to the SLAM problem in [15]. In SLAM, the update time and computational complexity of maintaining the coupled pose and scene covariance of the EKF algorithm scales quadratically $\boldsymbol{O(N^2)}$ with increasing the number of features $\boldsymbol{(N)}$. This in turn limits the number of feature matches available at any instant to that map so position accuracy is limited. SLAM with Compressed EKF (CEKF) was implemented in [16] by using a laser scanner in an outdoor environment to make computation cost constant until needed to update the whole map. The comparative performance evaluations of SLAM algorithms on the basis of position accuracy and processing time requirements were done in [17]. Among different versions of KF, the CEKF was more efficient for an outdoor environment where a lot of features are present. Inverse depth parametrization for features was proposed in [8] that can handle both close and distant features within the standard EKF framework. It remains well behaved for features at both stages (initialization and tracking) of SLAM processing, but has the drawback in computational terms that each point is represented by a 6D state vector as opposed to the 3D of a Euclidean XYZ representation. Dense scene flow combined with stereo V-SLAM was implemented in [18] and considerable positioning error was improved in moving objects. Nearby features give less linearization error on position estimation from V-SLAM so a camera centric coordinate system was proposed to reduce linearization error in [9].

In [19], Alcantarilla et al. proposed VO priors for robust EKF-SLAM. Constant linear and angular velocity model of the camera was replaced by VO prior. That performed quick pose estimation using the two stages RANSAC: a two-point algorithm for rotation followed by a one-point algorithm for translation. The result was better than constant velocity model in stereo vision model but very good prior information in the camera rotation was needed. Hwang and Song [20] used multiple features for stable monocular SLAM. Three types of features such as corners, line and illumination were used to make the landmark state vectors. That test was verified inside the room. The stable localization from multiple groups of features can be obtained but the arbitrary detection and association of different features on outdoor scenes is difficult to realize. Williams and Reid [21], proposed the V-SLAM system with persistent map by incorporating the additional information available from VO style measurements into the filter. A summary of their work was to produce a VO system using the monocular SLAM on each frame. Our objective, in different ways, is to run V-SLAM and VO separately and fusion of position information. V-SLAM runs in lower frequency than the VO so the computational complexity of maintaining the coupled pose and scene covariance become less than the related work. Hernandez et al. [22] used V-SLAM with partial odometry information from IMU by using inverse depth representation of features. That work was performed in a structured environment with uniform landmarks and the range within a meter.

3 Proposed approach

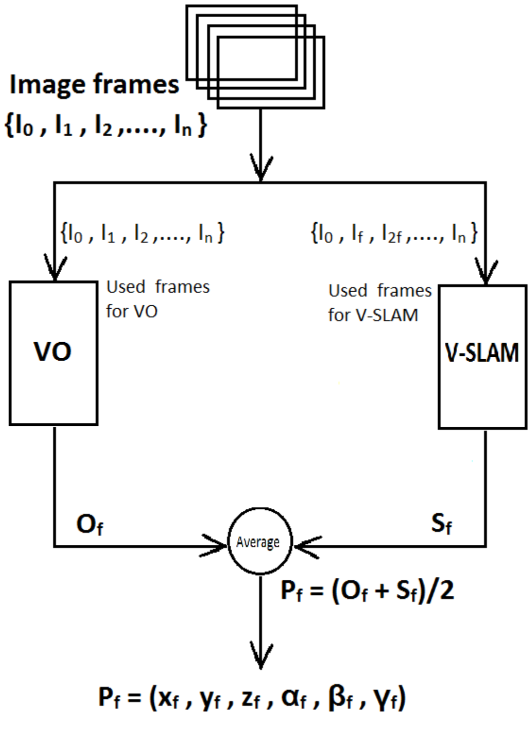
Image frames
$\{I_0, I_1, I_2, \ldots, I_n\}$
$\{I_0, I_1, I_2, \ldots, I_n\}$
$\{I_0, I_f, I_{2f}, \ldots, I_n\}$
Used frames for VO
Used frames for V-SLAM
VO
V-SLAM
O_f
S_f
Average
$P_f = (O_f + S_f)/2$
$P_f = (x_f, y_f, z_f, \alpha_f, \beta_f, \gamma_f)$

Fig. 1. Block diagram of visual odometry combined with fractional V-SLAM.

The block diagram of the proposed method is shown in Fig. 1. The camera works as the positioning sensor and the image sequences $\boldsymbol{I_{0:n} = \{I_0, I_1, I_2, ..., I_n\}}$ are the input of the system. The VO and V-SLAM are two subsystems of the proposed method and run respective algorithms to deduce position information of UAV from image sequences. VO uses every consecutive image frame to extract relative change in position from 2D-2D transformation. The estimated position from VO is notated as $\boldsymbol{O_f}$**.** V-SLAM uses interleaved image frames for the position estimation to reduce the computational burden and the estimated position is notated as $\boldsymbol{S_f}$**.** As the interleaved distance increases, the computational cost decreases but there is a need for a minimum number of matched features to estimate position. For this reason, we cannot go beyond the threshold in normal movement of the UAV. The estimated position from two subsystems is fused after every interleaved frame by simply averaged and notated as $\boldsymbol{P_f}$**.** This new value is used as reference to estimate the next position in both subsystems. Two statements on proposed approaches can be claimed based on hypotheses stated on related works. First one is uncertainty on position estimation in our work will be less than V-SLAM because of we have used the more number of features than in [6]. Second, the computational cost will be less in our approach than in [21] because of SLAM runs in interleaved input image sequences.

4 Experimental setup

The experiment is performed on data available from Library for Visual Odometry 2 (LIBVISO2)[1], Karlsruhe Institute of Technology. Reason of selecting this data is higher feature density, high feature matching speed-up of factor and it supports monocular ego motion estimation. Beside dataset those libraries contain camera calibration parameters of respective sequences so it makes it easy to test our approach in a long outdoor environment. 759 image frames of resolution 1344x391 pixels with Portable gray map (.pgm) are used for position estimation. A notebook computer with core i5 2.6-GHz processor, 8 GB memory and 64 bit operating system is used to perform position estimation from visual data. The proposed approach is processed in MATLAB R2011b. Corner features are detected by the FAST descriptor because of its short detection and matching time. Minimum number of landmarks used to perform visual position in this work is 25, which is more than sparse V-SLAM so this leads to more accurate position estimation. Interleaved distance 5 is used and weights of fusion are tuned fifty-fifty for VO and V-SLAM by empirical basis. Qualitative analysis of various weights on VO and V-SLAM are tested and among them average value gives better results.

Accuracy of estimated position is the main evaluating parameter of this work. The accuracy of the estimated position depends on the utility of the computation resource such as memory, processor and time. Therefore, computation resources are another evaluating parameter, which signifies how much computing resource that has been spent to achieve a certain accuracy level. Trajectory plot evaluates accuracy of estimated position in this work. The estimated position consists of six parameters of the UAV location information. The starting position of the UAV is origin with orientation aligned with respective axes. The accuracy of estimated position is observed by comparing with the approximated ground truth. Performing the test on the same data by varying one parameter and keeping all other fixes give relative comparisons. Relative comparison by adjusting the number of features is performed. We have compared trajectory evaluation and computing time evaluation of proposed approach with V-SLAM and VO without bundle adjustment. The comparison and the discussion are discussed in the following section.

5 Obtained results

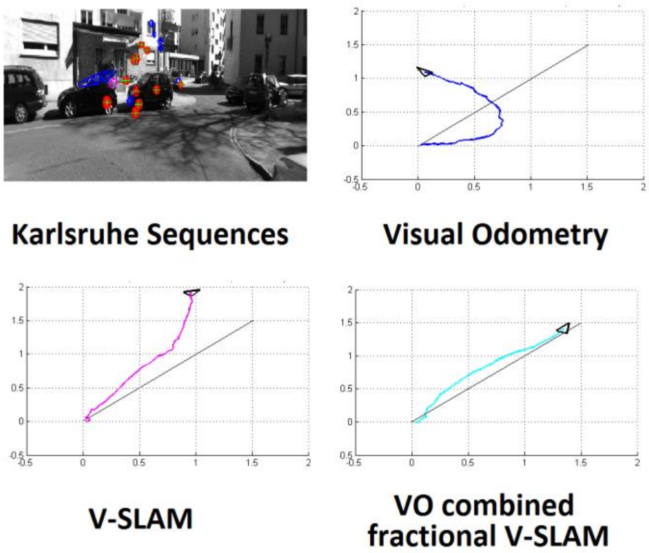

Fig. 2. Comparison of trajectory by three methods VO, V-SLAM and our approach using the same 750 image frames. The top left window shows datasets and the remaining three windows show trajectories by three different methods, both axes represent the distance and scaled to 100 meters on unit scale.

Fig. 2 shows the comparison of trajectory by three different methods. The experiment is performed on the Karlsruhe sequences. The top left window shows the dataset, the matched inliers features, matched outlier features and unmatched features are pointed by three different color red, magenta and blue respectively. The position estimation is performed on the basis of matched inliers points. Other three windows show the trajectory of the estimated position by three techniques. One unit of axes is scaled to a hundred meters. The triangle represents the estimated position of UAV. The black straight line shows the approximate ground truth. The camera moves hundreds of meters in the forward direction. The top right windows show VO trajectory, the bottom left shows V-SLAM trajectory and the bottom right shows our approach. Due to deficiency of optimization the estimated position starts to deviate after travelling around 75 meter far from origin. The orientation error of the camera greatly increases so the heading direction deviates in the wrong direction. Trajectory of the camera from V-SLAM is comparatively far better than VO. Due to the consideration of a limited number of features, counted as 25, the trajectory is rough. In our approach the trajectory is more smooth and straight compared to the rest of others. So from this result we can say that the accuracy of position estimation from our approach is superior to VO without pose optimization and V-SLAM with the same number of features.

Table 1. Computation time with respect to number of features

	Number of features	15	17	19	21	23	25
Computation time (sec)	VO	14	15	15	16	16	17
	V-SLAM	362	437	522	615	712	830
	Our approach	155	184	218	255	296	341

Table 1. shows the variation of computation time with respect to the number of features that are used to perform position estimation by visual odometry, V-SLAM and our approach. The comparison is done by using data set from Karlsruhe sequences of 1351x374 pixels and 200 frames. Based on the number of features used to perform position estimation the computing time is compared. The computation time to perform position estimation in VO does not increase significantly with varying the number of features from 15 to 25. But in V-SLAM the computation time varies from 362 sec to 830 sec if the number of features increases from 15 to 25. The rate of increasing shows nonlinearity relation between the computation time and the number of features. In these two experiments both VO and V-SLAM run in every sequence. In our approach the V-SLAM runs in a fractional manner so the computation time is less compared to V-SLAM. By minimizing the number of features the computation time by our method is considerably reduced. Our approach gives better computing time saving than V-SLAM with the same rate of varying the number of features. In summary, based on the results from Figure 2 and Table I, it can be said that our approach over performs than the individual VO and V-SLAM.

6 Conclusion

The key information from the above experiments is VO combined with V-SLAM works on outdoor environments for position estimation via single camera. The proposed approach on this work over performs than the individual VO without optimization and V-SLAM. The computing cost of the proposed approach is also less than V-SLAM, but there must be sufficient number of matched features in interleaved frames. Outdoor environments with less sharp corner features need different image features to handle visual positioning. Blur in the image frame and motion of the image feature creates a high number of unmatched features so decreasing correlation threshold can be a better alternative. There is a need for a different approach in visual position estimation if the large portion of scene is moving.

References

1. Shakhatreh, H., Sawalmeh, A.H., Al-Fuqaha, A., Dou, Z., Almaita, E., Khalil, I., Othman, N.S., Khreishah, A. and Guizani, M., 2019. Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges. IEEE Access, 7, pp.48572-48634.
2. Shen, C., Chang, T.H., Gong, J., Zeng, Y. and Zhang, R., 2020. Multi-UAV interference coordination via joint trajectory and power control. IEEE Transactions on Signal Processing.
3. Mostegel, C., Wendel, A. and Bischof, H., 2014, May. Active monocular localization: Towards autonomous monocular exploration for multirotor mavs. In 2014 IEEE international conference on robotics and automation (ICRA) (pp. 3848-3855). IEEE.
4. Ghosh, D. and Kaabouch, N., 2016. A survey on image mosaicing techniques. Journal of Visual Communication and Image Representation, 34, pp.1-11.
5. Aqel, M.O., Marhaban, M.H., Saripan, M.I. and Ismail, N.B., 2016. Review of visual odometry: types, approaches, challenges, and applications. SpringerPlus, 5(1), p.1897.
6. Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6):1052–1067, 2007.
7. David Niste´r, Oleg Naroditsky, and James Bergen. Visual odometry. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), volume 1, pages I–652. IEEE, 2004.
8. Javier Civera, Andrew J Davison, and J Montiel. Inverse depth parametrization for monocular slam. IEEE Transactions on Robotics, 24(5):932–945, 2008.
9. Jose´ A Castellanos, Jose´ Naira, and Juan D Tardo´s. Limits to the consistency of ekf-based slam 1. In 5th IFAC Symposium on Intelligent Autonomous Vehicles, 2004.
10. Javier Civera, Oscar G Grasa, Andrew J Davison, and JMM Montiel. 1-point ransac for extended kalman filtering: Application to real-time structure from motion and visual odometry. Journal of Field Robotics, 27(5):609–631, 2010.
11. Davide Scaramuzza and Friedrich Fraundorfer. Visual odometry [tuto- rial]. Robotics & Automation Magazine, IEEE, 18(4):80–92, 2011.
12. Davide Scaramuzza and Roland Siegwart. Appearance-guided monocu- lar omnidirectional visual odometry for outdoor ground vehicles. IEEE Transactions on Robotics, 24(5):1015–1026, 2008.
13. Davide Scaramuzza, Friedrich Fraundorfer, and Roland Siegwart. Real- time monocular visual odometry for on-road vehicles with 1-point ransac. In IEEE International Conference onRobotics and Automation (ICRA), pages 4293–4299. IEEE, 2009.
14. Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustmenta modern synthesis. In Vision algorithms: theory and practice, pages 298–372. Springer, 2000.

15. Randall Smith Matthew Self Peter Cheeseman, R Smith, and M Self. A stochastic map for uncertain spatial relationships. In 4th International Symposium on Robotic Research, pages 467–474, 1987.
16. John Folkesson and Henrik Christensen. Outdoor exploration and slam using a compressed filter. In ICRA, pages 419–426, 2003.
17. Gurkan Tuna, Kayhan Gulez, V Cagri Gungor, and T Veli Mumcu. Evaluations of different simultaneous localization and mapping (slam) algorithms. In IECON 2012-38th Annual Conference on IEEE Industrial Electronics Society, pages 2693–2698. IEEE, 2012.
18. Pablo Fernа´ndez Alcantarilla, Jose´ J Yebes, Javier Almaza´n, and Luis Miguel Bergasa. On combining visual slam and dense scene flow to increase the robustness of localization and mapping in dynamic environments. In 2012 IEEE International Conference on Robotics and Automation (ICRA), pages 1290–1297. IEEE, 2012.
19. Pablo Ferna´ndez Alcantarilla, Luis Miguel Bergasa, and Frank Dellaert. Visual odometry priors for robust ekf-slam. In 2010 IEEE International Conference onRobotics and Automation (ICRA), pages 3501–3506. IEEE, 2010.
20. Seo-Yeon Hwang and Jae-Bok Song. Monocular vision-based slam in indoor environment using corner, lamp, and door features from upward-looking camera. Industrial Electronics, IEEE Transactions on, 58(10):4804–4812, 2011.
21. Brian Williams and Ian Reid. On combining visual slam and visual odometry. In 2010 IEEE International Conference on Robotics and Automation (ICRA), pages 3494–3500. IEEE, 2010.
22. E Hernandez, Juan Manuel Ibarra, Jose´ Neira, R Cisneros, and JE Lavin. Visual slam with oriented landmarks and partial odometry. In 2011 21st International Conference on Electrical Communications and Computers (CONIELECOMP), pages 39–45. IEEE, 2011.