A Method of Target Tracking Based on Monocular Vision for Mobile Robot in Unknown Environment

Wu Ming*, Guangyuan Fu, Hongqiao Wang, Shuai Huang
Research Institute of High Technology Xi’an, P.R. China

*Corresponding author e-mail: hyacinth531@163.com

Abstract. In order to solve the problem of target tracking based on monocular vision for mobile robot in unknown environment, a robot simultaneous localization, map building and target tracking method is proposed. This method can realize the simultaneous online estimation of robot, environment features and target states by using the bearings only observations from monocular sensors. In the process of estimation, the monocular vision based Simultaneous localization and mapping (MVSLAM) runs independent of the Object Tracking (OT), as the basis of the OT, MVSLAM provides the state information of the robot platform for the OT, and OT uses the bearing measurements of target and robot state to estimates the state of target. MVSLAM is based on inverse depth parameterization and running under framework of full probability Extended Kalman Filter (EKF), and another EKF is also used for target tracking independently. The simulation experiments show the performance of the design method, and analyze the problem of observability of Bearing-only tracking. To solve the observability problem, a robot motion control method is presented, under the control of this method, the robot will make satellite surround motion around the target. Finally, the feasibility and accuracy of the method are verified by simulation experiments.

1. Introduction (Heading 1)

The application of Simultaneous Localization, Mapping and Object Tracking (SLAMOT) can help to realize the simultaneous online estimates states of robot, environment features and target. The research of this problem enables robot to have real-time cognition ability to understand situation of external environment and state of itself, so has wide application prospects, such as domestic robot [1], Security robot [2] and so on. Most of the researches on SLAMOT have been based on active sensor [3, 4] or stereoscopic vision sensor [5]. However, those kinds of sensors have defects such as large volume, heavy weight, high energy consumption, high cost of equipment and limited observation distance, which limit the practical application range of SLAMOT. As a passive sensor, monocular vision sensor is becoming more and more attractive in robotics because of its small volume, low energy consumption, low cost, high detail presentation and easy to understand. This paper is to study the SLAMOT problem based on monocular vision.

Monocular vision based target tracking from mobile robot in unknown environment involves many hot topics in communities of computer vision and intelligent mobile robots, including Structure from Motion (SFM), Visual SLAM (VSLAM), Bearing Only Target Motion Analysis (BOTMA). The key points in these research fields are different. SFM focuses on the use of image sequences to achieve 3D
modelling and estimation of mobile platform trajectory [6]. This method usually assumes that the environment is static, and the moving target is treated as environmental noise [7], and does not achieve estimation of moving target state. STF based on the multiple view geometry method to calculate relative pose of robot in discrete time series, there is a problem of accumulative error in pose calculating. To solve this problem, bundle adjustment [8] of batch optimization is often used to correct errors at regular intervals, thus reducing the real-time and operational efficiency of the system. V-SLAM focuses on the use of monocular or stereoscopic vision to solve the problem of simultaneous on-line estimation of the state of the robot and the environment. The operation efficiency of the related methods [9] is high, but when dealing with the environment target, it also treats target as environment noise, and the target state is not estimated [10]. The objective of BOTMA is to obtain state estimation of a moving target by utilizing a sequence of bearing measurements, received by a mobile platform or fixed stations at different locations. BOTMA methods usually assume that the state of the observation platform is known, so the state of the platform itself and the running environment is no need to estimate. The relative research of BOTMA focuses on problem of observability [11] and algorithm of estimation [12].

In this paper, a simultaneous localization, map building and object tracking algorithm based on monocular vision is proposed. The algorithm uses monocular visual SLAM and target tracking as two independent and interdependent modules. Independence means that the monocular vision SLAM independently uses the extended Kalman filter based on inversion depth to estimate the state of the robot and the environment, and the target state does not like the method proposed in paper[3] expanded with the states of robot and the environments to form the system state, however, on the basis of the monocular vision SLAM, the state of robot by VSLAM and the measurement of target by sensor are used to estimate the state of the target under the extended kalman filter framework. Interdependence refers to the fact that the robot needs to maneuver on the basis of the target state to ensure the accuracy of the target state estimation, since an observability problem of the target tracking caused by the bearing only tracking. Therefore, there is a coupling relationship between the robot control and the target state estimation, some researchers call at “Dual Effect” [13].

The paper is organized as follows. Section 2 formulates the estimation problem, gives states the assumptions made throughout the paper, and proposes robot and target motion model and target observation model. Section 3, provides the basic framework of the algorithm, designs the estimation method based on Kalman filtering, focuses on the generation method of observation residual matrix, which is key point to target state update. The simulation results of target tracking without robot maneuvering control and an analysis of the observability problems are given in section 4, also developing a maneuvering control method to make robot move around the target, the simulation verifies the feasibility and accuracy of the control method to target state estimation. The conclusion and extension of the current work to the problem is discussed in Section 5.

2. Model Formulation
Monocular vision based target tracking for mobile robots in unknown environments involves simultaneous iterative estimation of different object states. The sensors used include internal odometry and external monocular camera. They provide data of robot state update and observation of external environment. Due to slippage and turbulence, there is accumulative error in the odometer, and in the process of projecting the 3D environment to the 2D image, the monocular camera lost depth information. Key points to solving in this paper are the coupling estimation of the related objects and to overcome the defect of the sensor observation. Assuming the process follows Markov chain. Next, we will introduce motion and observation models involved in the algorithm.

2.1. Target Motion Models
The Constant Velocity Model [14] is used to describe kinematics of target. The state of the target at the time \( t \) is composed of the spatial Cartesian coordinates \((x_t', y_t', z_t')\) and the velocity components
\((x'_t, y'_t, z'_t)\) in the direction of X, Y and Z axis. That is \(X'_t = [x'_t, y'_t, z'_t, \dot{x}'_t, \dot{y}'_t, \dot{z}'_t]^T\). The linear discrete transition function of target state is as follows:

\[
X^t_k = f^t(X^t_{k-1}, A^t_{k|k-1}, q^t_{k|k-1}) = A^t_{k|k-1} \cdot X^t_{k-1} + q^t_{k|k-1}
\]  

(1)

And \(A^t_{k|k-1} = \exp(F \cdot \Delta t)\), \(\Delta t\) is the sampling interval, \(q^t_{k|k-1}\) is Gauss white noise with mean value 0, and the covariance matrix of \(q^t_{k|k-1}\) is:

\[
Q^t_{k|k-1} = \int_0^{\Delta t} \exp(F \cdot (\Delta t - \tau)) \cdot L \cdot Q^t \cdot (L^T) \cdot \exp(F \cdot (\Delta t - \tau)) \cdot d\tau
\]  

(2)

2.2. Robot Motion Model

At the time \(k\), the state of robot \(X^r_k\) is composed by position component \(p^r_k = [x^r_k, y^r_k, z^r_k]^T\) and attitude component \(q^r_k = [q_{1k}, q_{2k}, q_{3k}, q_{4k}]^T\), robot motion model define as follows:

\[
X^r_k = f(X^r_{k-1}, u^r_{k-1})
\]  

(3)

\(u^r_{k-1}\) is control input or odometer measurements vectors, it form is:

\[
u^r_{k-1} = [\Delta p^r_k, \Delta e^r_k] = [\Delta x^r_k, \Delta y^r_k, \Delta z^r_k, \Delta \phi^r_k, \Delta \theta^r_k, \Delta \psi^r_k]^T
\]  

(4)

\(\Delta p^r_k\) is position increment of robot, \(\Delta e^r_k\) is attitude increment for robot, expressed as three Euler angle. Specifically, the update model of robot position state component \(p^r_k\) is:

\[
p^r_{k+1} = p^r_k + R^r_k \cdot \Delta p^r_k
\]  

(5)

\(R^r_k\) is rotation matrix calculated by \(k\) time robot attitude component. \(\Delta p^r_k = [\Delta x^r_k, \Delta y^r_k, \Delta z^r_k]^T\) is the robot position increment obtained by the odometer.

The update model of robot attitude state component \(q^r_k\) is:

\[
q^r_{k+1} = q^r_k + \frac{1}{2} \cdot \Omega(\Delta e^r_k) \cdot q^r_k
\]  

(6)

\(\Omega(\Delta e^r_k)\) is skew symmetric matrix.

2.3. Target observation model

The target observation model realizes the mapping from the robot state \(X^r_k\), the target state \(X^t_k\), the pose of the camera in the robot coordinate system \(X^R_{k,c}\) to the image observation value \(z^t_k\). It gives as follows:
\[
\mathbf{z}_k^1 = h^1(X_k^1, X_k^r, X_k^{R,c}, d, s) + r
\]

\[s = [x_u, y_v, u_0, v_0]^T\] is camera internal parameter vector, \(d = [d_1, d_2]^T\) is distortion coefficient of camera. \(r\) is Gauss white noise.

The transformation process of observation is shown in Figure 1.

Firstly, the transformation of target global coordinate state \(X_k^{W,t}\) to the target state in robot coordinate \(X_k^{R,t}\) is realized by using function \(X_k^{R,t} = f^{W\rightarrow R}(X_k^{W,t}, X_k^{W,R})\). Secondly, by using the pose of camera in robot coordinate system, the coordinate transformation of \(X_k^{R,t}\) to \(X_k^{C,t}\) is conducted by \(X_k^{C,t} = f^{R\rightarrow C}(X_k^{R,t}, X_k^{R,C})\). Thirdly, by using the small aperture imaging model, the transformation from camera coordinate \(X_k^{C,t}\) to 2-D coordinate of image plane \(I, t\) is realized by \(x_k^{I,t} = f^{C\rightarrow I}(X_k^{C,t})\). Fourthly, using tangential distortion coefficient \(d\) to generate coordinates change \(x_k^{ID,t}\) by \(x_k^{ID,t} = f^{R\rightarrow ID}(x_k^{R,t}, d)\). Finally, transforming coordinates after distortion \(x_k^{ID,t}\) to pixel plane coordinates \(z_k^t\) by \(z_k^t = f^{ID\rightarrow P}(x_k^{ID,t}, s)\).

### 3. Algorithm framework and solution method

#### 3.1. Algorithm framework

The target tracking method of mobile robot in unknown environment is carried out by EKF framework. The overall process is shown in Figure 2.
First, the control components (including angular velocity $\Delta \phi$ and linear velocity $\Delta x$) are generated according to the estimated state of robot and target state at $k$ time. Second, the robot uses method introduced in section 1 to carry out the SLAM based on monocular vision and gets the estimated state $r_k^X$ and covariance $r_k^P$ of the robot. Third, the target predicts its own state $t_{k-1}^X$ and covariance matrix $t_{k-1}^P$ according to the motion model. Fourth, using $r_k^X$ and $t_{k-1}^X$ to predict target observation $z_{k-1}^z$, and to generate the observed residual matrix $S_{k-1}^1$. Finally, the target state $t_k^X$ and variance matrix $t_k^P$ are updated using the actual observed values $z_k^z$ of the current target.

### 3.2. V-SLAM

The paper uses the inverted depth monocular vision SLAM method proposed in document [15], the processing flow is shown in Figure 3. The algorithm uses the fully associative extended Kalman filter framework, and the system state is extended by the state of the robot and states of environment features.

---

**Figure 2.** General flow chart of target tracking method

**Figure 3.** Flow chart of monocular SLAM
In the prediction stage, the robot's motion model and environmental observation model are used to predict the robot's state and covariance matrix. In observation data association stage, the correspondence between the observed values of the environmental features and the known environmental features of the system is determined by the feature point matching method. In EKF update phase, the system state and covariance matrix are updated using observed residuals. In environmental feature management stage, it mainly completes the addition of new environment features, the deletion of old environment features, and the conversion of depth feature vectors to Cartesian coordinate vectors.

3.3. Bearing only target tracking method

3.3.1. Target state and covariance prediction. The prediction formula of target is as follows:

\[
X_t^i = A^i_{k|k-1} \cdot X_{k-1}^i \tag{8}
\]

\[
P_t^i = A^i_{k|k-1} \cdot X_{k-1}^i \cdot A^i_{k|k-1}' + Q \tag{9}
\]

\(A^i_{k|k-1}\) is State transition matrix of target Constant Velocity Model, \(Q\) is error matrix.

3.3.2. Prediction of target observation and observation residual matrix. The target predicted observations \(z_t^i\) is obtained by using the target observation model (7) Using error propagation formula to generate target observation residual matrix as:

\[
S_t^i = H^{t\rightarrow z} \cdot P_t^i \cdot H^{t\rightarrow z}' + H^{t\rightarrow z} \cdot P_t^i \cdot H^{c\rightarrow z} + H^{c\rightarrow z} \cdot P_c^i \cdot H^{c\rightarrow z}' + H^{s\rightarrow z} \cdot P_s^i \cdot H^{s\rightarrow z}' + H^{d\rightarrow z} \cdot P_d^i \cdot H^{d\rightarrow z}' + R \tag{10}
\]

Where, 
\[
H^{t\rightarrow z} = \left. \frac{\partial h^i}{\partial X_k^i} \right|_{x|k-1} \cdot X_k^i, \quad H^{t\rightarrow z} = \left. \frac{\partial h^i}{\partial X_k^c} \right|_{x|k-1} \cdot X_k^c, \quad H^{c\rightarrow z} = \left. \frac{\partial h^i}{\partial S_k} \right|_{x|k-1} \cdot S_k, \quad H^{s\rightarrow z} = \left. \frac{\partial h^i}{\partial d} \right|_{x|k-1},
\]

\(H^{d\rightarrow z} = \left. \frac{\partial h^i}{\partial d} \right|_{x|k-1}\) are Jacobian matrix of \(h^i\) to \(X_k^i\), \(X_k^c\), \(X_k^c\), \(S\), \(d\) individually.

3.3.3. Target state, covariance update. According to the extended Kalman filtering principle, the updated target state \(X_k^i\) and covariance matrix \(P_k^i\) are as follows:

\[
X_k^i = X_k^i + K \cdot (z_k^i - z_k^i) \tag{11}
\]

\[
P_k^i = P_k^i - K \cdot S_k^i \cdot K' \tag{12}
\]

\[
K = P_k^i \cdot H_k^{t\rightarrow z} \cdot (S_k^t)^{-1} \tag{13}
\]

\(K\) is Kalman gain matrix.
4. Experiment and results
The experiment is carried out in the Matlab2017 environment. The scene of simulation is shown in Figure 4.

The range experimental scene is: X=50~50(m), Y=50~ -50(m), Z=10~10(m). Among them, the feature points of the environment are evenly distributed. The state of robot is represented by the triangle, in which the green triangle is the estimated state, the blue triangle is the actual state. The actual state of the target is represented by the red solid point, and the target estimation state is represented by the green ellipse.

The initial state of the target is: \( \mathbf{x}_0 = [0 \ 40 \ 10 \ -3.5 \ 0] \), the initial state of the robot is \( \mathbf{r}_0 = [20 \ 40 \ 10 \ 0 \ 0] \). Trajectory of the target is pure linear, the motion process of target does not disturbed any noise, and EKF filter of OT adopts Constant Velocity Model, and the uncertainty coefficient of the target motion is 0.7. Figure 5 shows the overall tracking results.

![Figure 4. Simulation experiment scene](image1)

![Figure 5. Overall tracking results of method](image2)
The real lines in the graph are actual trajectories of target and robot. The dotted lines are estimated trajectories of target and robot. The cross points are locations of the environmental features, and ellipses are the estimated location distribution of environment features.

The target always keeps moving in a straight line and eventually stops at location of $(0, -12.5)$, and the robot follow the target movement stops at location of $(0, -3.6)$. From the target estimation curve can see that the error of the target estimation in the direction of the X axis is small, but the error in the direction of the Y axis is huge leading to the great deviation of the final target estimation position and the actual position. From the robot estimation curve can see that the system can always estimate the robot state accurately.

Analyzing the results is that in the front 100s, the robot makes a turning movement, and this kind of motion leading monocular vision sensor produced enough parallax to help the robot achieved target depth estimation. However, after 100s, the robot begins to follow the target and does not have any changed on the direction of the line of sight, so the depth information of the target cannot be updated all the time, which eventually leads to the deterioration of the actual and estimated state of the target.

Another robot control method is used to make the satellite movement around the target. This motion only needs to set the Y component of position control to be a constant value. Figure 6 shows the overall tracking results.

![Figure 6](image_url)

**Figure 6.** Overall tracking results of satellite around control

Figure 7 is error curve of robot and target estimation, and comparison of the actual and estimated trajectories using robot satellite around control method.
Fig. 7. (a) And (b) are estimation error curves of robot and target at different time and the contrast diagram between the actual and the estimated trajectory. The four sub-figures on the left side are X, Y, Z state components estimated errors with different time. From the error curves, the robot can always keep a good estimation accuracy of its position, and the average error of position estimation is 0.1016 (m). In contrast, the tracking accuracy of the target is lower and the average error of the position estimation is 1.3022 (m). However, the tracking accuracy of the maneuver is significantly improved compared to the results of the robot's non-maneuvering motion.

5. Conclusion

This paper proposes a method of simultaneous localization, mapping and object tracking by using the bearings only sensor. Monocular vision based Simultaneous localization and mapping section runs independent of the Object Tracking section, and establishing connection to object tracking by measurement of target. Experimental results show that robot maneuver plays an important role in bearing only target tracking. Next, we will focus on robot optimal control method to achieve accurate estimation of the target in unknown environment.

Acknowledgments

This work was supported by the National Natural Science Foundation for Young Scientists of China (Grant No: 61503389), and National Science Basic Research Plan in Shanxi Province of China (Grant No: 2016JM6061).

References

[1] C. Zhao; L. Sun; R. Stolkin. A fully end-to-end deep learning approach for real-time simultaneous 3D reconstruction and material recognition. Proceedings of 18th International Conference on
Advanced Robotics (ICAR), Hong Kong, China, 2017, pp.343 - 351.

[2] R. Salman, I. Willms, T. Sakamoto. Environmental imaging with a mobile UWB security robot for indoor localisation and positioning applications. In Proc of European Microwave Conference, Nuremberg, Germany, 2013, pp. 1643 - 1646.

[3] C. C. Wang, C. Thorpe. Simultaneous localization and mapping with detection and tracking of moving objects. International Journal of Robotics Research, vol.16, no.9, pp. 889-916, 2007.

[4] M. Wu, J. Y. Sun. Extended kalman filter based moving object tracking by mobile robot in unknown environment. Robot, vol.32, no.3, pp. 334 – 343, 2010.

[5] A. Ess, B. Leibe, K. Schindler, L. Gool, A mobile vision system for robust multi-person tracking, Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2008, pp.1 - 8.

[6] G. Klein, D. Murray, Parallel tracking and mapping for small AR workspaces. Proceedings of 6th IEEE International Symposium on Mixed and Augmented Reality, 2007, pp.225 - 234.

[7] C. M. Cheng, X.A. Hsu, S.H. Lai, A novel structure from motion strategy for refining depth map estimation and multi-view synthesis in 3DTV, Proceedings of 2010 IEEE International Conference on Multimedia and Expo, 2010, pp.944 - 949.

[8] B. Triggs, P. Mclauchlan, R. Hartley, Bundle adjustment - a modern synthesis, Lecture Notes in Computer Science, vol.1883, pp.298 - 372, 2000.

[9] J. Civera, O. G. Grasa, 1-Point RANSAC for EKF filtering. application to real-time structure from motion and visual odometry, Journal of Field Robotics, vol.27, no.5, pp. 609 - 631, 2010.

[10] D. P. Zou, P. Tan, CosSLAM: Collaborative visual SLAM in dynamic environments, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.35, no.2, pp. 354-366, 2013.

[11] M. H. Ferdowsi, Observability conditions for target states with bearing-only measurements in three-dimensional case, Proceedings of the 2006 IEEE International Conference on Control Applications, 2006, pp.1444 - 1449.

[12] L. Ince, B. Sezen, E. Saridogan, An evolutionary computing approach for the target motion analysis (TMA) problem for underwater tracks. An International Journal of Expert Systems with Applications, vol.36, no.2, pp.3866 - 3879, 2009.

[13] J. Kim, S. Rock. Stochastic feedback controller design considering the dual effect. Aiaa Journal, 2013.

[14] B. Y. Shalom, X. R. Li. Multitarget-multisensor tracking: principles and techniques, USA, YBS Publishing, 1995.

[15] J. Civera, A. J. Davison, J.M.M. Montiel. Inverse Depth Parametrization for Monocular SLAM. IEEE Transactions on Robotics, vol.24, no.5, pp.932 - 945, 2008.