Vision-Based Target Localization with Cooperative UAVs Towards Indoor Surveillance

Guanchong Niu\textsuperscript{1}, Qi Cao\textsuperscript{1}, Chung Shue Chen\textsuperscript{2}
\textsuperscript{1}Guangzhou Institute of Technology, Xidian University, Guangzhou, China
\textsuperscript{2}Nokia Bell Labs, Paris-Saclay, 91300 Massy, France
Email: \{niuganchong, caoqi\}@xidian.edu.cn, chung_shue.chen@nokia-bell-labs.com

Abstract—Unmanned aerial vehicles (UAVs) have been widely adopted for a variety of civilian and military applications. Despite their many advantages, most UAVs are not suitable for indoor missions due to the lack of global position system (GPS) and the large size of UAV with diverse sensors. In this work, a lightweight and cost-effective indoor multi-UAV surveillance system is presented for accurate target localization. The proposed system employs a vision-based architecture, leveraging ORB-SLAM for self-localization and YOLOv3 for object detection. The triangulation method is employed for target positioning, offering reliable performance with negligible time delays and improved detection compared to depth camera approaches. After that, the Riccati observer is introduced to address the stochastic nature of the system. Through a series of experiments involving real UAVs, the system demonstrates its ability to accurately localize and track targets in indoor environments, updating the UAVs’ positions in real-time with optimal performance.

Index Terms—Cooperative UAV Systems, Indoor Surveillance, Vision-based Localization, Riccati Observer.

I. INTRODUCTION

The emergence of unmanned aerial vehicles (UAVs) is of great significance for promoting various civilian applications such as disaster response, routing patrolling, and express delivery. By taking advantage of the high mobility and low cost, UAVs have been widely employed in surveillance for public safety [1]. Despite the abundant studies on UAVs, there are very few applications that focus on indoor scenarios. Specifically, it is challenging to realize the multi-UAV localization and tracking for the indoor scenario due to the lack of global positioning system (GPS). To fill the gap, various localization algorithms are proposed, including ultra-wide band (UWB) [2], inertial sensor [3], visible light [4], and magnetic field [5]. However, the systems mentioned above either require additional sensors or exhibit poor performance. Besides, as an universal sensing technique in the indoor environment, WiFi-based tracking systems for indoor services have attracted a great interest in recent years. Main techniques can be classified into four categories, namely received signal strength indication (RSSI)-based triangulation [6], fingerprint mapping [7], angle of arrival (AoA) measurement, and time of flight (ToF) measurement. Unfortunately, the accuracy of existing WiFi-based systems relies on the deployment of access point (APs) and dramatically decreased by the non-line-of-sight (NLoS) transmission.

Besides the above WiFi-based methods, several studies have developed multi-UAV surveillance systems based on additional sensors, such as radar, radio frequency (RF), acoustic and video methods. For instance, a localization algorithm was derived in [8], where a UAV is equipped with a monostatic multiple-input-multiple-output (MIMO) radar for target detection and localization. In [9], the light-detection-and-ranging radar (LiDAR), as a high-accuracy radar, was employed on the UAV for target-ranging measurement. In addition, RF-based surveillance was deployed using the direction-of-arrival (DoA) estimation to localize an object by analyzing the characteristics of the amplitude and phase differences of incoming signals. However, the aforementioned systems require additional sensors, which may result in unaffordable load and cost. Moreover, the UAV equipped with multiple sensors may be too large to be employed for the indoor scheme due to safety concerns.

Using the camera-only UAVs, several studies have developed the multi-UAV surveillance systems for diverse tasks. For instance, a detection model was proposed for action recognition to support searching and rescue (SAR) after natural disaster in [10]. The detection model was deployed on-board to automatically execute long-range real-time SAR mission. Furthermore, [11] considered the UAVs equipped with thermal cameras for landmine detection. The aforementioned studies focused on outdoor applications, while the indoor surveillance systems are still an open problem. Some pioneering works emerged in recent years. In [12], an indoor and outdoor UAV surveillance application was presented for industrial management using various sensors including LiDAR, ultrasound and camera. In [13], a deep learning-based navigation algorithm of UAVs was proposed for indoor surveillance, which requires large training overhead. The implementation of indoor schemes is still limited due to safety concerns and low accuracy, making it difficult to expand its applications. To cope with the challenges of indoor surveillance, target detection and localization are urgently required for vision-based UAV systems.

In summary, the main contributions of this paper are elaborated as follows:
- **Real-World Implementation:** A multi-UAV cooperative surveillance system has been fully implemented with all necessary software and hardware. In order to validate the effectiveness of the proposed system, both simulation and practical experiments were conducted for multi-UAV localization and tracking. The numerical results obtained have demonstrated that our proposed system outperforms the radar-empowered and RGB-D (Red Green Blue-Depth) cameras in terms of performance.

- **Algorithm Design and Analysis:** The proposed localization algorithm that uses monocular vision achieves accuracy comparable to that of sophisticated RGB-D cameras. Target positioning is achieved through a triangulation method that operates in two steps: prediction and update. To tackle the stochastic nature of the system, Riccati observers are proposed, known for their effectiveness in dealing with stochastic systems.

Notation: Uppercase boldface and lowercase boldface letters are used to denote matrices and vectors, respectively. $I_N$ denotes the identity matrix with size $N \times N$. $\|A\|$ stands for the $L^2$ norm of $A$. Note that $\hat{A}$ is the estimator of $A$, where $\hat{A}$ can be a matrix or vector. $A^T$ and $A^H$ denote the transpose and conjugate transpose of $A$, respectively. Besides, $\dot{A}$ represents the derivative of $A$.

### II. System Model

The hardware design of a multi-UAV indoor system must be lightweight and cost-effective in order to ensure safe and efficient operation. The architecture of the proposed vision-based indoor surveillance systems is illustrated in Fig. 1, where multiple quadrotor UAVs with a server and router are deployed to recognize and localize the target. Server, such as a computer, keeps receiving the RGB images from UAVs and sending commands to UAVs. The UAVs with a single camera take videos and execute instructions.

#### A. Dynamic Model

We consider the scenario that $N$ UAVs are deployed to localize a moving target. All UAVs can communicate with the server via a router. The position and velocity of the moving target are referred to $p^t \in \mathbb{R}^3$ and $v^t \in \mathbb{R}^3$, i.e.,

$$\dot{p}^t = v^t, \quad (1)$$

Correspondingly, the position and velocity of the moving target are referred to $p^i \in \mathbb{R}^3$ and $v^i \in \mathbb{R}^3$. As we have no knowledge of the target position, the dynamic model of the target is formulated using the following equation:

$$x^i_k(k + 1) = A_r x^i_k(k) + \delta_r(k), \quad (2)$$

where $x^i_k = [p^{iT}, v^{iT}]$, $A_r = \begin{bmatrix} I_3 & \Delta T I_3 \\ 0 & I_3 \end{bmatrix}$, and $\delta_r(k) \in \mathcal{N}(0, Q)$ is Gaussian noise with a covariance matrix $Q$. It is worth noting that we assume the velocity of the target is a constant.

The problem we are focusing on in this study is to estimate the state of the moving target using the positions and velocities of the UAVs. The main challenge here is the stochastic nature of the system due to the presence of Gaussian process noise $\delta_r$. To tackle this problem, we propose to use Riccati observers which are known for their effectiveness in dealing with stochastic systems. The Riccati observers will use the available measurements from the UAVs and the dynamic model of the target to provide an estimate of the target’s state. The detailed description of the Riccati observers and how they are implemented in this scenario will be discussed in the subsequent sections.

#### B. Hardware Design and Self-Localization Approach

For the sake of safety concerns, the size of UAVs is limited to be less than 200 grams. To reduce the cost, none of additional sensor is required to be installed in the indoor environment. Therefore, camera is the only on-board sensor on UAVs. As illustrated in Fig. 1, the server communicates with UAVs via a router to receive image information and send commands.

SLAM is a method commonly applied on mobile platforms for self-exploration in various environments. The unique feature of SLAM is that the process of mapping and localization is performed recursively and simultaneously. ORB-SLAM [14] can process sequences from indoor and outdoor scenes and from car, robot, and hand-held motions. Its accuracy is typically below 1 cm in small indoor scenarios and of a few meters in large outdoor scenarios. ORB-SLAM is commonly considered as the most reliable and complete solution to monocular SLAM. It achieves unprecedented performance with respect to other state-of-the-art monocular SLAM approaches. Therefore, ORB-SLAM is employed in this work. Moreover, object detection is realized by YOLOv3 [15].

![Fig. 1. Schematic diagram for the proposed vision-based indoor surveillance systems.](image)
III. TARGET POSITIONING WITH MONOCULAR VISION

The current research on UAV-based surveillance simply obtains the position of the target in the field of view. The joint monitoring with multiple UAVs involves the problem of trajectory planning, which requires accurate world coordinate information with targets. Low-latency, accurate target positions can be easily obtained with a depth camera. Here, we propose a method that achieves similar performance as the depth camera. Fig. 2 shows the overall design, which basically contains three parts.

Fig. 2. Framework of the proposed target positioning method.

A. Target Detection and Position Estimation

As discussed in Section II-B, ORB-SLAM is used to perform the UAV positioning mission. The main challenge is how to make the coordinates obtained by each UAV in the same coordinate system.

Each UAV node contains an independent ORB-SLAM process. The camera poses generated by ORB-SLAM are referred to as the initial key frames, which means that the pose matrices are described in those frames. As a result, each process has its initial key frame. In our design, one UAV performs SLAM while storing a pre-scanning map whose size is unlimited, as long as it contains the initial key frame. The UAV nodes will then share the map among themselves to execute the positioning tasks.

Fig. 3 shows the frames with suitable world coordinates \{W\} established according to the actual situation. Denote by \( T_f \) the pose matrix of initial key frame described in the camera frames \{C_1\}, \{C_2\}, \cdots, \{C_N\}. According to the transformation relationship in Fig. 3, the desired pose matrix described in frame \{W\} can be expressed as:

\[
P_w^w = \left( T_f^c \times T_f^w \right)^{-1}, \tag{3}
\]

where \( P_w^w \in \mathbb{R}^{4\times4} \), \( n = 1, \cdots, N \) are the pose matrices of UAVs, respectively.

Furthermore, \( T_f^w \in \mathbb{R}^{4\times4} \) is the transformation matrix between the initial key frame and the world frame, and is given by:

\[
T_f^w = \begin{bmatrix} R_f^w & t_f^w \end{bmatrix},
\]

where \( R_f^w \in \mathbb{R}^{3\times3} \) is the rotation of frame (\( \text{Rot} \)) around the X, Y, Z axes by \( \alpha_f, \beta_f, \gamma_f \) in turn (by intrinsic rotation) while \( t_{w_{org}}^w \in \mathbb{R}^{3\times1} \) is the original point of world frame described in the above frame.

The positioning module is embedded into the UAV node, which is run as an independent process to keep processing input images and generating the locations of UAVs.

B. Scaled Triangulation

Given the relation \( \lambda_i l_i = T_i e \) for each observed point \( l_i \) and its corresponding camera matrix \( T_i \), we want to estimate the real-world position \( e \) using the triangulation method. Since \( e \) is a homogeneous 4D coordinate, we can write it as \( e = [p ; w] ^\top \), where \( p \in \mathbb{R}^{3\times1} \) is the 3D real-world coordinate and \( w \) is the homogeneous scale factor.

To construct a linear system, we can rearrange the relation as:

\[
\lambda_i l_i = T_i [p ; w] ^\top. \tag{6}
\]

To eliminate the scale factor \( \lambda_i \), we can form a linear equation based on the cross product property:

\[
l_i \times (T_i [p ; w] ^\top) = 0. \tag{7}
\]

Expanding the cross product, we thus get a system of linear equations for each \( l_i \):

\[
\begin{bmatrix} l_i^x \\ l_i^y \\ l_i^z T_i \end{bmatrix} \begin{bmatrix} p \\ w \end{bmatrix} = 0. \tag{8}
\]

Here, \( l_i^x \) represents the skew-symmetric matrix form of \( l_i \). By stacking all the equations, we form a linear system:

\[
Ae = 0 \tag{9}
\]

where \( A = [l_1^x T_1, \cdots, l_N^x T_N]^T \).

To solve the homogeneous system, we can perform singular value decomposition (SVD) on the matrix \( A = USV^\top \).

The solution \( e \) is the eigenvector corresponding to the smallest singular value in \( S \). The estimated real-world
point $p$ can be reconstructed by normalizing $\epsilon$ by its last coordinate (since the solution is up to a scale):
\[
\hat{p} = \frac{1}{w} \left[ \hat{p}_x \hat{p}_y \hat{p}_z \right]
\]

where $\hat{p}_x$, $\hat{p}_y$, and $\hat{p}_z$ are the first three coordinates of the estimated real-world position.

IV. LOCALIZATION USING RICCATI OBSERVERS

A. Riccati Observers

The optimal estimation problem for the state of the moving target can be solved using a Riccati observer. To formulate the problem, we consider the observation model of the UAVs. Suppose that at each time step $k$, the $i$-th UAV measures the position of the target with a measurement noise $\eta_i(k) \in \mathcal{N}(0, \mathcal{R})$. Then, the observation model can be written as:

\[
y_i(k) = H \hat{x}^i(k) + \eta_i(k), \quad i = 1, \cdots, N,
\]

where $y_i(k)$ is the measurement of the $i$-th UAV, $H = [I_3 \ 0]$ is the measurement matrix, and $\eta_i(k)$ is the Gaussian measurement noise with a covariance matrix $\mathcal{R}$.

The problem then is to estimate the state of the target $\hat{x}^i(k)$, given the measurements $y_i(k)$, the system dynamics, and the observation model. The optimal estimate, in the sense of minimizing the mean squared error, can be obtained using a Riccati observer.

The Riccati observer, also known as the Kalman filter, is a recursive estimator that operates in two steps: prediction and update. In the prediction step, the observer predicts the state of the target at the next time step based on the current state estimate and the system dynamics. In the update step, the observer corrects the predicted state using the measurement from the UAVs.

The Riccati observer equations are:

\[
\hat{x}^i(k|k-1) = A_r \hat{x}^i(k-1|k-1),
\]

\[
\hat{x}^i(k|k) = \hat{x}^i(k|k-1) + K_k(y_i(k) - H \hat{x}^i(k|k-1)),
\]

where $\hat{x}^i(k|k-1)$ is the predicted state, $\hat{x}^i(k|k)$ is the updated state, and $K_k$ is the Kalman gain which is computed using the Riccati equation. The initial state estimate $\hat{x}^i(0|0)$ is usually set to a known initial condition or a zero vector.

The optimization problem is aiming to find the Riccati observer that minimizes the estimation error covariance. This problem can be solved using the Riccati equation, which is a matrix differential equation. The Riccati equation provides a recursive formula to compute the optimal Kalman gain $K_k$ and the error covariance matrix at each time step.

B. Continuous Riccati Equation

In this part, we consider the solution for Riccati observation optimization. We define the error covariance matrix $P(t) = E[(x^t - \hat{x}^t)(x^t - \hat{x}^t)^T]$, where $\hat{x}^t$ is the state estimate.

The continuous-time Riccati equation for the error covariance matrix $P(t)$ is given by:

\[
\dot{P}(t) = A_r P(t) + P(t) A_r^T - P(t) H^T \mathcal{R}^{-1} H P(t) + Q,
\]

where $\mathcal{R}$ and $Q$ are the measurement noise covariance matrix. The optimal Kalman gain $K(t)$ is then given by:

\[
K(t) = P(t) H^T \mathcal{R}^{-1}.
\]

Finally, the state estimate is updated using the Kalman gain:

\[
\dot{x}^t = A_r \hat{x}^t + K(t) (y_i - H \hat{x}^t).
\]

By integrating the equations, we can obtain the optimal state estimate and the error covariance matrix at any given time point. It should be emphasized that the primary benefit of employing a continuous-time Riccati equation lies in delivering a more precise estimation of the system state, particularly when confronted with rapidly evolving system dynamics.

V. EXPERIMENTAL RESULTS

A. Experimental Setup

The UAV utilized in the proposed system’s experiments is the Tello, manufactured by DJI [16]. The Tello was chosen for its extensive Application Programming Interfaces (APIs), cost-effectiveness, stable performance, and most critically, its ultra-light weight. Comprehensive information about the Tello can be located in the referenced citation [16].

B. General Comparison of Target Positioning by Different Methods

To compare the target positioning accuracy achieved by different methods, the following experiment has been performed. The RGB-D camera is located at (-4650 mm, 0 mm, 1400 mm), UAV 1 is located at (-4650 mm, 700 mm, 1700 mm) whereas UAV 2 is located at (-4650 mm, -700 mm, 1700 mm). The target moves along the Y-axis at a constant speed.

The result is shown in Fig. 4. The sub-figures from the top to bottom show the change of X, Y and Z coordinates over the time. Furthermore, the small black and green dots represent the original data estimated by the triangulation and the data from RGB-D camera, respectively. Furthermore, the red line is the processed data by the Kalman filter while the blue line is the target position predicted by the predictor. For presentation clarity, a zoomed-in image taken from 40000 mm to 60000 mm is shown. On the other hand, Fig. 5 gives the error comparison. By analyzing Fig. 4
1000 2000 3000 4000 5000 6000 7000 8000 9000 10000 11000
Disatance (mm)
0
20
40
60
80
100
120
Localization Error (mm)
Error for the Proposed Methods
Error for RGB-D Camera

Fig. 4. Comparison of target positioning performance (zoom in).

Fig. 5. Error comparison for different systems.

and Fig. 5, the following interesting observations can be obtained:

1) The overall trend of all data is consistent with the depth camera data, indicating that the entire positioning system has no systematic errors;
2) Notably, a discernible shift is observed between the red line and the green dots, indicating a time delay of approximately 500 ms. However, this time delay is not manifested in the predicted target position when compared to the RGB-D data;
3) The deviation of the green dot increases with the distance between the RGB-D camera and the target. Furthermore, the detection performance of the depth camera is observed to be negatively affected by distance, and the effective detection distance is smaller than the proposed method.

C. Experiment of Tracking Target

This experiment aims to assess the monitoring and tracking capabilities of the proposed DSS. The target is set to move randomly within an indoor setting, while two UAVs are programmed to maintain a distance of 4000 mm from the target for tracking purposes. As illustrated in Fig. 6, each UAV’s viewpoint frames the target upon detection. The coordinate system depicted on the right represents the layout of the room, upon which the position of the target (green dot), the positions of UAV 1 (red dot with a short line) and UAV 2 (blue dot with a short line), as well as the desired positions of UAV 1 (small dark red dot with a short line) and UAV 2 (small dark blue dot with a short line), are updated in real time. As soon as the tracking process is initiated, the Graphical User Interface (GUI) deploys the UAVs to their new positions based on the real-time positions of the UAVs and the target. Subsequently, the controller module executes the necessary actions to maneuver the UAVs to their planned locations.

The set distance from the target is specified as 4000 mm. During the course of this experiment, the actual distance between the UAVs and the target is computed, as shown in Fig. 7. The red line illustrates the real distance between UAV 1 and the target, while the blue line indicates the distance between UAV 2 and the target. The black line, on the other hand, represents the desired distance. Fig. 7 provides statistical data on the distance between the UAVs and the target. The average and standard deviation of the distance errors are determined to be acceptable for reliable monitoring and tracking of moving targets.

VI. Conclusion

In conclusion, this paper developed and validated a lightweight, cost-effective indoor multi-UAV surveillance
system for accurate target localization. The proposed vision-based architecture effectively integrates ORB-SLAM for self-localization and YOLOv3 for object detection, ensuring reliable performance in various indoor environments. The dynamic modeling and Riccati observers employed in this work demonstrate the ability to cope with the stochastic nature of the system, resulting in accurate localization of moving targets. A series of experiments involving Tello drones provided strong evidence that the proposed system outperforms depth camera-based approaches in terms of localization accuracy and detection performance over increasing distances. Furthermore, the real-time monitoring and tracking capabilities of the system demonstrated its applicability in practical surveillance scenarios.

REFERENCES


